

# Estadística aplicada uni y multivariante

Andrés Muñoz Serrano

$$S_{YX}^2 = \frac{1}{n} \sum (X_i - \bar{X})^2$$

16	0.844	0.855	0.874	0.889	0.901
17	0.851	0.863	0.887	0.895	0.905
18	0.858	0.869	0.892	0.906	0.914
19	0.863	0.874	0.897	0.910	0.917
20	0.868	0.879	0.901	0.914	0.919
21	0.873	0.884	0.905	0.917	0.920
22	0.878	0.888	0.908	0.920	0.923
23	0.881	0.892	0.911	0.923	0.926
24	0.884	0.895	0.914	0.926	0.929
25	0.888	0.898	0.916	0.928	0.930
26	0.891	0.901	0.918	0.930	0.931
27	0.894	0.904	0.920	0.933	0.935
28	0.896	0.906	0.923	0.935	0.936
29	0.898	0.908	0.924	0.936	0.937
30	0.900	0.910	0.926	0.937	0.939
31	0.902	0.912	0.927	0.937	0.939
32	0.904	0.914	0.927	0.939	0.940



9930 14504

va que la cantidad de cabras va aumentando. Como el valor de  $\hat{Y}$  es un número real del año (podríamos haber puesto, sencillamente, desde el ordenada en el origen nos da la cantidad estimada de cabras en España

o de cabras estimado para el año próximo es

$$\hat{Y}_{1989} = 0.05760 \times 1989 = 3661.4$$

## Consejería de Agricultura y Pesca

co de esta predicción es

$$S_{YX} = \sqrt{S_{YX}^2} = \sqrt{\frac{1}{n} \sum (X_i - \bar{X})^2}$$

lo que se puede realizar pruebas de hipótesis para contrastar la hipótesis nula de  $\hat{Y}$  es una estima de  $\mu$  por medio de  $\bar{X}$

Row	Col	Pct	Frequency
no	enfermo	11	11
no	noenfer	22.00	22.00
no	Total	32.35	32.35
si	enfermo	39.00	39.00
si	noenfer	74.00	74.00
si	Total	67.65	67.65
Total	enfermo	34.00	34.00
Total	noenfer	27.00	27.00
Total	Total	54.00	54.00
Total	enfermo	40.91	40.91
Total	noenfer	66.00	66.00
Total	Total	106.91	106.91



Tomo II

# **ESTADÍSTICA APLICADA UNI Y MULTIVARIANTE**

**TOMO I**

Andrés Muñoz Serrano  
Departamento de Genética  
Universidad de Córdoba

**ESTADÍSTICA APLICADA UNI Y MULTIVARIANTE (2 volúmenes). Tomo I**

© *Edita:* JUNTA DE ANDALUCÍA. **Consejería de Agricultura y Pesca.**

*Publica:* Viceconsejería. Servicio de Publicaciones y Divulgación.

*Autor:* Muñoz Serrano, Andrés

*I.S.B.N.:* 84-8474-070-6 (Tomos I y II, obra completa)

*Depósito Legal:* SE. 3.996 (Obra completa)

*Fotocomposición e impresión:* J. de Haro Artes Gráficas, S.L. Parque Ind. P.I.S.A.  
Mairena del Aljarafe • Sevilla

<b>CAPÍTULO 1</b>	<b>Introducción</b>	<b>1</b>
	Estadística	1
	Estadística y cuantificación de la incertidumbre	2
	La Estadística y el Método Científico	6
<b>CAPÍTULO 2</b>	<b>Fundamentos</b>	<b>11</b>
	Distribuciones	11
	Función de probabilidad	11
	Función de distribución	12
	Experimento, espacio muestral y suceso	12
	Ejemplo	12
	Ejemplo	13
	Distribuciones de tipo discreto	13
	Distribuciones continuas	14
	Distribuciones como modelos para representar situaciones reales	18
	Variables	19
	Clases de variables por su escala de medida	19
	Variable <i>cuantitativa</i> / variable <i>cualitativa</i>	21
	Variable <i>discreta</i> / variable <i>continua</i>	21
	Datos de respuestas categóricas	22
	Variable de <i>Respuesta</i> y variable <i>Explicativa</i>	22
	Exactitud y precisión de los datos	22
	Ejemplo	23
	Población	24
	Muestra	24
	Parámetro de una población	24
	Estadísticos muestrales y estimas	25
	Parámetros de posición o de localización y de dispersión	25
	Media	25
	Moda	26
	Mediana	26
	Media geométrica	26
	Media armónica	26
	Varianza	27
	Desviación típica	28
	Coefficiente de variación	29

Consideraciones a tener en cuenta en la elección del análisis apropiado y en la elección de los parámetros apropiados a las variables que se están estudiando	29
Parámetros adecuados para describir las variables nominales	30
Parámetros adecuados para describir las variables ordinales	30
Parámetros adecuados para describir las variables de intervalo	31
Parámetros adecuados para describir las variables de proporción	31
Análisis apropiado para los diferentes tipos de variables	31
Análisis Univariantes/Multivariante	34
Tipos de análisis	34
Pruebas de homogeneidad	35
Regresión	35
Correlación	36
Análisis de Componentes Principales y Análisis Factorial	37
Problemas de Clasificación	37
Datos categóricos	38
<b>CAPÍTULO 3</b>	
	<b>Pruebas de hipótesis</b>
Error Tipo I, Error Tipo II	43
Función de potencia de una hipótesis nula	44
Estadístico de prueba y regla de decisión	45
Diferentes tipos de Hipótesis	53
Pruebas de hipótesis de dos colas	54
Sistematización del método para realizar una prueba de hipótesis	58
Prueba de hipótesis sobre la media de una distribución normal	59
Ejemplo	64
Prueba de hipótesis sobre la varianza de una distribución normal	65
Ejemplo	66
Error Tipo I, Error Tipo II y tamaño de muestra	67
Ejemplo	70
Ejemplo	71
Elección del tamaño de muestra	73
Prueba de hipótesis de la $p$ de una distribución binomial aproximándola a la normal	75
Ejemplo	78
Prueba de hipótesis de la $\lambda$ de una distribución de <i>Poisson</i> aproximándola a la normal	79
Ejemplo	80

**CAPÍTULO 4**

	<b>Una/dos poblaciones:</b>	83
	<b><math>t</math> de Student y <math>T^2</math> de Hotelling</b>	84
	Ideas elementales sobre los Modelos	85
	El Modelo lineal aditivo	87
	Análisis de la Varianza (ANOVA)	90
	Ejemplo	92
	Supuestos del modelo sin factores	94
	Parámetros estimados en el modelo sin factores	95
	Prueba de hipótesis en el modelo sin factores; $t$ de Student	97
	Ejemplo	97
	C4-1.SAS	99
	Ejemplo	99
	Generalización a más de una variable	100
	La $T^2$ de Hotelling	100
	Ejemplo	101
	C4-2.SAS	103
	Dos Poblaciones	
	Comparación de las varianzas de dos poblaciones normales	104
	Ejemplo	107
	Comparación de las medias de dos poblaciones normales	110
	Supuestos paramétricos y limitaciones de la prueba $t$	113
	Ejemplo	114
	C4-3.SAS	115
	Dos poblaciones con análisis de varianza	115
	Ejemplo	116
	C4-4.SAS	116
	Generalización multivariante	117
	Ejemplo	117
	C4-5.SAS	
	Comparación de medias muestrales usando observaciones emparejadas	119
	Ejemplo	121
	C4-6.SAS	124
	Modelo factorial sin repetición con un solo grado de libertad para los tratamientos	125
	C4-7.SAS	125
	Confusión de modelos	126
	C4-8.SAS	126
	Generalización multivariante	127
	Ejemplo	128
	C4-9.SAS	
	Caso especial: un solo ejemplar comparado con una muestra	129
	Comparación de dos medias de muestras independientes y varianzas desiguales	130
	Aproximación de Cochran y Cox	132

Comparación de medias cuando $\sigma_1^2 \neq \sigma_2^2$ y $n_1 = n_2 = n$	132
Comparación de medias cuando $\sigma_1^2 \neq \sigma_2^2$ y $n_1 \neq n_2$	133
Aproximación de <i>Satterthwaite</i>	134
Ejemplo	134
C4-10.SAS	136
Importancia de los supuestos paramétricos	137
Ejemplo	137
C4-11.SAS	138
Elección del tamaño de muestra en comparaciones de dos o más muestras	139
Ejemplo	143
Ejemplo	145
Comparación de dos proporciones binomiales	147
Ejemplo	149

## CAPÍTULO 5

	<b>Análisis de la Varianza y Análisis Multivariante de la Varianza</b>	
Generalización		149
Ideas elementales sobre los Modelos		150
El Modelo lineal aditivo		150
Factor, tratamientos y efectos		154
Planteamiento del Análisis de la Varianza		158
Ejemplo		159
Comparación de Tratamientos		161
Ejemplo		162
Análisis de varianza de una Vía o del Diseño Completamente Aleatorio		164
Modelo del análisis de Varianza de una Vía		164
Descomposición de la Varianza en un Modelo de una Vía		164
Descomposición de los grados de libertad en un Modelo de una Vía		166
Parámetros estimados en el Modelo de una vía		166
Prueba de Hipótesis en el modelo de una vía		168
Ejemplo		169
C5-1.SAS		170
Ejemplo		171
C5-2.SAS		172
Generalización del ANOVA de una vía a más de una variable		173
Ejemplo		173
C5-3.SAS		174
Es suficiente considerar un solo factor		175
Ejemplo		175
Descomposición de la varianza en un modelo factorial sin repetición		178
Descomposición de los grados de libertad en un modelo factorial sin repetición		179

Parámetros estimados en un Modelo Factorial sin repetición	180
Prueba de hipótesis en un modelo factorial sin repetición	180
Ejemplo	181
C5-4.SAS	183
Generalización del ANOVA de dos vía a más de una variable	184
Ejemplo	184
C5-5.SAS	184
Generalización del método para el análisis de modelos factoriales sin repetición con más de dos factores	185
Ejemplo	188
C5-6.SAS	190
Inconvenientes del cuadrado latino	191
Ejecución del diseño.	191
Ejemplo.	192
C5-7.SAS	193
Diseños conmutativos o cruzados o sobrecruzados o con intercambio ( <i>cross-over</i> )	193
Ejemplo	196
C5-8.SAS	197

<b>CAPÍTULO 6</b>	<b>Supuestos paramétricos y Transformaciones</b>	201
	Supuestos del análisis de varianza	201
	Aleatorización.	201
	Aditividad de los tratamientos y de los efectos ambientales	203
	Prueba de no aditividad.	204
	Ejemplo.	205
	Normalidad del error	206
	Pruebas de normalidad con una sola muestra	207
	Ejemplo	207
	Prueba de <i>Kolmogorov-Smirnov</i> para una muestra	208
	Ejemplo	208
	Prueba de <i>Shapiro Wilk</i> para varias muestras	209
	Ejemplo	210
	C6-1.SAS	211
	Independencia de los errores	212
	Pruebas de rachas	213
	Prueba de rachas para variables binomiales	214
	Ejemplo	215
	Pruebas de rachas para datos cuantitativos	215
	Ejemplo	216
	Coefficiente de correlación serial	216
	Ejemplo	217
	C6-2.SAS	218
	Homogeneidad de varianzas	219
	Prueba de <i>Bartlett</i>	220
	Ejemplo	221



	C6-3.SAS	222
	Transformaciones	223
	Transformación logarítmica	223
	Ejemplo	224
	C6-4.SAS	224
	Transformación de la raíz cuadrada	226
	Ejemplo	226
	C6-5.SAS	227
	Transformación angular o arcoseno	229
	Ejemplo	230
	C6-6.SAS	231
<b>CAPÍTULO 7</b>		
	<b>Modelos Factoriales</b>	235
	Número de medidas por casilla	235
	Ejemplo	236
	Interacción	239
	Factores de <i>efectos fijos</i> y factores de <i>efectos aleatorios</i>	241
	Modelo de efectos fijos	244
	Modelo de efectos aleatorios	245
	Modelo mixto	246
	Análisis de varianza Factorial con repetición	247
	Modelo Factorial con repetición	247
	Descomposición de la varianza y de los grados de libertad en un modelo factorial con repetición	248
	Parámetros estimados en un Modelo Factorial con repetición	249
	Pruebas de hipótesis en un modelo factorial con repetición	254
	Ejemplo	261
	C7-1.SAS	265
	Experimentos factoriales	266
	Ejemplo	267
	Generalización del método para el análisis de modelos factoriales con repetición con más de dos factores	268
	Ejemplo	271
	C7-2.SAS	271
	Un factorial con repetición en un factorial sin repetición	272
	Ejemplo	272
	C7-3.SAS	274
<b>CAPÍTULO 8</b>		
	<b>Modelos Jerárquicos</b>	277
	Relación entre los factores	277
	Ejemplo	277
	Diseños Jerárquicos o con subgrupos	282
	Ejecución del diseño	284
	Modelo lineal	284
	Descomposición de la varianza en un modelo jerárquico	284
	Descomposición de los grados de libertad en un modelo jerárquico	285

	Parámetros estimados en un modelo jerárquico	286
	Pruebas de hipótesis en un modelo jerárquico	290
	Ejemplo	294
	C8-1.SAS	297
	Generalización del método para el análisis de modelos jerárquicos con más de dos factores	299
	Ejemplo	301
	C8-2.SAS	304
<b>CAPÍTULO 9</b>	<b>Modelos Compuestos</b>	<b>307</b>
	Ejemplo	307
	Más sobre las interacciones	308
	Generalización del método para el análisis de Modelos Compuestos	310
	Modelo compuesto con un jerárquico dentro de una interacción	310
	Modelo compuesto con una interacción de un factor principal con un factor jerárquico	312
	Cálculo de las <i>estimaciones</i> de los <i>CM</i> por el método rápido.	313
	Ejemplo	316
	C9-1.SAS	320
	Ejemplo	321
	Prueba <i>F</i> aproximada	324
	Ejemplo.	326
	C9-2.SAS	329
	Estima conjunta ( <i>englobe</i> ) de la significación de varios factores en modelos compuestos.	331
	Ejemplo	333
<b>CAPÍTULO 10</b>	<b>Comparación de Medias</b>	<b>337</b>
	Comparaciones o pruebas planeadas	338
	Ejemplo	344
	C10-1.SAS	346
	Modelos con más de un factor	347
	Ejemplo	347
	C10-2.SAS	348
	Ejemplo	349
	C10-3.SAS	350
	Contrastes ortogonales con factores cuantitativos	352
	Ejemplo	353
	C10-4.SAS	354
	Contrastes ortogonales con modelos factoriales de dos factores, ambos cuantitativos	355
	Descomposición de la suma de cuadrados de la interacción	356
	Ejemplo	358
	C10-5.SAS	362
	Experimentos para el estudio de las <i>Superficies de Respuesta</i>	363

Ejemplo	364
C10-6.SAS	365
C10-7.SAS	368
C10-8.SAS	370
Comparación múltiple de medias o contrastes a posteriori o pruebas sugeridas por los datos.	372
Prueba <i>LSD</i> (Least-Significant-Difference)	373
Prueba <i>Tukey</i> o de recorrido Studentizado	373
Prueba de <i>Student-Newmans-Keuls</i> o <i>SNK</i>	374
Prueba de <i>Duncan</i> o de amplitudes múltiples	375
Prueba de Scheffe	375
Ejemplo	376
Prueba <i>Duncan</i>	377
Prueba Student-Newman-Keuls ( <i>SNK</i> )	378
Prueba <i>Tukey</i>	380
Prueba <i>LSD</i> o de mínima diferencia significativa	381
Prueba Scheffe	382
C10-9.SAS	384
C10-10.SAS	386
Ejemplo	388
C10-11.SAS	389

**CAPÍTULO 11**

	<b>Regresión</b>	391
Introducción		391
Relación funcional/Relación causal		393
Regresión lineal		393
Estima de la recta de regresión por mínimos cuadrados		396
Modelo y supuestos de la regresión		399
Estima de la recta de regresión		402
Ejemplo		404
Fuentes de variación en la regresión.		405
Prueba de ajuste		408
Coefficiente de determinación y coeficiente de alineación o factor de mejoramiento		409
Ejemplo		411
Pruebas de hipótesis e intervalos de confianza para el coeficiente de regresión		412
Ejemplo		413
Ejemplo		414
Pruebas de hipótesis e intervalos de confianza de la estimación, por interpolación, de un valor <i>Y</i>		414
Ejemplo		415
Ejemplo		415
Pruebas de hipótesis e intervalos de confianza para la media de la variable dependiente		416
Ejemplo		416
Ejemplo		417

Pruebas de hipótesis e intervalos de confianza de la ordenada en el origen	418
Ejemplo	418
Ejemplo	419
Análisis de los residuo ( $e_{Y,X}$ )	419
Ejemplo	420
Análisis de las <i>influencias</i>	421
Residuos <i>studentizados</i>	421
Diferencias de ajustes	422
<i>D</i> de Cook	422
Medida tipificada de lo extrema que es una observación	423
Proporción de las varianzas generalizadas	423
Diferencias tipificadas de las <i>b</i> y de las <i>a</i>	423
Ejemplo	424
C11-1.SAS	427
C11-2.SAS	429
Varios valores de <i>Y</i> por cada valor de <i>X</i>	430
Ejemplo	433
C11.3.SAS	434
Valores de regresión y valores ajustados	436
Prueba de homogeneidad de dos o más líneas de regresión	437
Ejemplo	438
C11-4.SAS	440

**CAPÍTULO 12**

**Predicciones y Ajustes**

Pruebas de hipótesis e intervalos de confianza de la predicción, por extrapolación, de un valor <i>Y</i>	443
Ejemplo	444
Ejemplo	445
C12-1.SAS	446
Predicción de <i>X</i> a partir de <i>Y</i> . Calibración lineal	447
Ejemplo	448
Ajuste de la recta por el origen	449
Ejemplo	450
C12-2.SAS	453
C12-3.SAS	455
Transformación <i>Probit</i> en los casos de predicción de <i>X</i> a partir de una <i>Y</i> no normal	456
Ejemplo	459
C12-4.SAS	460
C12-5.SAS	463
El modelo logístico o LOGIT	464
C12-6.SAS	466
Modelo II de regresión o cuando la <i>X</i> se mide con error	467
Recta mínimos rectángulos	469
Pruebas de hipótesis e intervalos de confianza con la recta mínimos rectángulos	470

Ejemplo	471
C12-7.SAS	473
Recta de los tres grupos de <i>Bartlett</i>	475
Ejemplo	476
C12-8.SAS	478
<b>CAPÍTULO 13</b>	
	<b>Regresión Múltiple</b>
Dos variables independientes	481
Ejemplo	481
Prueba de ajuste en la regresión múltiple	483
Coefficiente de determinación y coeficiente de alineación o factor de mejoramiento	484
Ejemplo	486
Pruebas de hipótesis e Intervalos de confianza para los coeficientes de regresión	487
Ejemplo	489
Pruebas de hipótesis e intervalos de confianza para la ordenada en el origen	490
Ejemplo	491
C13-1.SAS	492
Pruebas de hipótesis e intervalos de confianza para la estimación por interpolación	493
Ejemplo	495
Pruebas de hipótesis e intervalos de confianza de la predicción, por extrapolación, de un valor $Y$	496
Ejemplo	497
Pruebas de hipótesis e intervalos de confianza para la media de la variable dependiente	498
Ejemplo	499
C13-2.SAS	500
Análisis de los residuo ( $e_{y,x}$ )	501
C13-3.SAS	502
Análisis de las <i>influencias</i>	505
Residuos <i>stundetizados</i> ( $R_{student}$ )	505
Diferencias de ajustes ( $D_{ffits}$ )	505
$D$ de <i>Cook</i> ( $Cook.s D$ )	506
Medida tipificada de lo extrema que es una observación ( $Hat\ Diag\ H$ )	506
Proporción de las varianzas generalizadas ( $Cov\ Ratio$ )	506
Diferencias tipificadas de las $b$ y de la $a$ (... $D_{fbetas}$ , $INTERCEP\ D_{fbetas}$ )	507
C13-4.SAS	507
Valores de regresión y valores ajustados	509
Ejemplo	510
C13-5.SAS	511
Prueba de homogeneidad de dos o más líneas de regresión	514
Ejemplo	515

	C13-6.SAS	515
	Importancia relativa de diferentes variables $X$	518
	Selección del modelo adecuado	519
	Ejemplo	521
	C13-7.SAS	522
<b>CAPÍTULO 14</b>	<b>Regresión Curvilínea</b>	<b>533</b>
	Ajustes de curvas	533
	Regresión no lineal	533
	Curvas logarítmicas	535
	Curvas de crecimiento exponencial	535
	Ejemplo	536
	C14-1.SAS	537
	C14-2.SAS	538
	Crecimiento alométrico	538
	Ejemplo	539
	C14-3.SAS	540
	C14-4.SAS	541
	Curvas polinómicas	542
	Ejemplo	543
	C14-5.SAS	545
	C14-6.SAS	549
	Polinomios ortogonales	552
	Ejemplo	554
	Cómo separar la suma de cuadrados	555
	Ejemplo	556
	C14-7.SAS	557
<b>CAPÍTULO 15</b>	<b>Correlación simple</b>	<b>561</b>
	Correlación y regresión	561
	Función de densidad bivalente y supuestos paramétricos	562
	Coefficiente de correlación	565
	Coefficiente de alineación o factor de mejoramiento	566
	Relación entre los coeficientes de correlación y regresión	567
	Pruebas de hipótesis e intervalos de confianza para $\rho=0$ y tamaños de muestra superior a 50	569
	Ejemplo	570
	C15-1.SAS	571
	Pruebas de hipótesis para $\rho \neq 0$ y para $\rho = 0$ en muestras pequeñas	572
	Ejemplo	575
	Ejemplo	577
	Homogeneidad de dos coeficientes de correlación	578
	Ejemplo	579
	Ejemplo	580
	Homogeneidad de varios coeficientes de correlación	581
	Ejemplo	581

	Correlación intraclass o repetibilidad	582
	Ejemplo	584
	Pruebas de asociación no paramétricas	585
	Coeficiente de correlación de rangos de <i>Spearman</i>	585
	Coeficiente de correlación de clasificación de <i>Kendall</i>	586
	Ejemplo	587
	C15-2.SAS	589
	Coeficiente de correlación serial	590
<b>CAPÍTULO 16</b>	<b>Correlación multivariante</b>	591
	Correlación parcial	591
	Correlación parcial o correlación de los residuos	592
	Ejemplo	592
	C16-1.SAS	592
	Cálculo de la correlación parcial	593
	Ejemplo	595
	C16-2.SAS	595
	Correlación múltiple	597
	Ejemplo	598
	C16-3.SAS	599
	Ejemplo hipotético	601
	Conceptos básicos de la correlación canónica	602
	Primera correlación canónica	603
	Segunda (y sucesivas) correlación canónica	604
	Pruebas de hipótesis	606
	Representación de los valores de las variables canónicas	606
	C16-4.SAS	607
	Correlación canónica usando los componentes principales	609
	C16-5.SAS	610
	Aplicación al análisis discriminante	612
	Ejemplo	612
	C16-6.SAS	613
<b>CAPÍTULO 17</b>	<b>Análisis de la Covarianza</b>	619
	Usos del análisis de covarianza	619
	Control del error	620
	Ajuste de medias de tratamientos	620
	Interpretación de datos	621
	Partición de la covarianza	622
	Comparación de regresiones	622
	El modelo lineal del análisis de covarianza	622
	Ejemplo	625
	Supuestos paramétricos del análisis de covarianza	629
	Análisis de covarianza del modelo de una vía o del diseño completamente aleatorio	630
	Ajuste de las medias de tratamientos	633
	Pruebas de hipótesis	635

Aumento de precisión debido a la covarianza	635
Ejemplo	636
C17-1.SAS	641
Análisis de covarianza del modelo factorial sin repetición o del diseño de bloques aleatorios	643
Ajustes de las medias de los tratamientos y de los bloques	648
Pruebas de hipótesis	649
Aumento de precisión debido a la covarianza	651
Ejemplo	651
C17-2.SAS	660
Análisis de covarianza de un modelo factorial con repetición	663
Ajustes de las medias	668
Pruebas de hipótesis	670
Aumento de precisión debido a la covarianza	672
Ejemplo	673
C17-3.SAS	685
Análisis de covarianza de un modelo jerárquico o con subgrupos	689
Ajustes de las medias	693
Pruebas de hipótesis	695
Aumento de precisión debido a la covarianza	696
Ejemplo	697
C17-4.SAS	707
Generalización del método para la realización del análisis de covarianza	711
Partición de la covarianza	712
Ejemplo	714
Prueba de homogeneidad de coeficientes de regresión	717
Ejemplo	719
C17-5.SAS	721
Análisis de covarianza con varias covariables	722
Ejemplo	722
C17-6.SAS	722
Análisis multivariante de la covarianza. Varias variable dependientes con varias covariables	725
Ejemplo	725
C17-7.SAS	726

**CAPÍTULO 18**

**Análisis de Componentes Principales y Análisis Factorial**

	731
Interpretación geométrica	732
Ejemplo ilustrativo	733
Cálculo matemático	737
Número de componentes principales a retener	738
Interpretación de las componentes	739
Ejemplo	740
C18-1.SAS	740



Usando variables tipificadas	745
Cálculo matemático para estimar los valores y vectores propios de los datos tipificados	748
Número de componentes principales a retener	749
Interpretación de las componentes	750
Representación conjunta de individuos y variables	750
Ejemplo	751
C18-2.SAS	751
Datos sin tipificar o datos tipificados	755
ANÁLISIS FACTORIAL	756
Ejemplo ilustrativo	757
Conceptos básicos del Análisis Factorial	758
Método del Factor Inicial: Análisis de Componentes Principales	760
Método del Factor Inicial: Componentes Principales Iterados	763
Método de máxima verosimilitud	766
Rotación de los factores	767
Rotación varimax	768
Rotación oblicua	770
Asignando puntuaciones de los factores a los individuos	772
Ejemplo	774
C18-3.SAS	774
C18-4.SAS	777
Análisis de <i>tamaño y forma</i>	780
Ejemplo	781
C18-5.SAS	782
Detección de datos anómalos	787
Ejemplo	787
C18-6.SAS	787
C18-7.SAS	791
C18-8.SAS	795
Análisis de Varianza de las Componentes Principales	795
Ejemplo	796
C18-9.SAS	796
C18-10.SAS	801
Ejemplo	804
C18-11.SAS	805
C18-12.SAS	808
Componentes Principales de variables cualitativas	811
Ejemplo	812
C18-13.SAS	813
<b>CAPÍTULO 19</b>	
<b>Análisis discriminante y Cluster</b>	<b>823</b>
Análisis Discriminante	823
Ejemplo ilustrativo	824
Conceptos básico de clasificación	825
Interpretación de los coeficientes	828

Base teórica	828
Analogía con la regresión	829
Cálculo de la <i>función discriminante de Fisher</i>	831
Coefficientes tipificados	832
Probabilidades a posteriori	832
Incorporando las probabilidades a priori en la elección del punto <i>C</i>	833
Bondad de la función discriminante	835
Prueba de la contribución de las variables de clasificación	836
Ejemplo	836
C19-1.SAS	836
Análisis canónico discriminante	842
Ejemplo	844
C19-2.SAS	844
Selección de variables en el Análisis Discriminante	849
Ejemplo	850
C19-3.SAS	851
Análisis Cluster	853
Ejemplo ilustrativo	854
Conceptos básicos previos a la presentación de los análisis cluster: Diagrama de dispersión	854
Medidas de distancias	855
Técnicas analíticas de agrupamientos	857
Clasificación jerárquica	857
Método de las distancias mínimas	858
C19-4.SAS	859
Método del centroide	861
C19-5.SAS	862
Ventajas e inconvenientes de los procedimientos jerárquicos	863
Ejemplo	863
C19-6.SAS	863
Clasificación no jerárquica	866
Método de las <i>K-medias</i>	866
Ejemplo	866
C19-7.SAS	867
Ejemplo	868
C19-8.SAS	868
Métodos no paramétricos	872
Ejemplo	873
C19-9.SAS	873

**CAPÍTULO 20**

**Datos de Respuesta Categórica 881**

Clasificación	882
Uso de las pruebas de datos categóricos	883
Análisis de datos categóricos cuando se tiene una muestra	883
Prueba de bondad de ajuste	883

Realización de la prueba de bondad de ajuste. Distribución binomial	884
Ejemplo	884
Realización de la prueba de bondad de ajuste. Distribución multinomial	886
Ejemplo	886
Grados de libertad en las pruebas de bondad de ajuste	887
Ejemplo	887
Fórmula general para el cálculo del estadístico $\chi^2$	888
Prueba de razón de verosimilitudes	889
Ejemplo	890
Corrección de continuidad	890
Ejemplo	891
Clasificaciones de una vía o clasificación única	892
Ejemplo	892
Otras pruebas no paramétricas para una sola variable de clasificación	893
Prueba de <i>Kolmogorov-Smirnov</i> para una muestra	893
Prueba de <i>Shapiro Wilk</i> para varias muestras	893
Independencia de los errores	893
Pruebas de rachas	893
Prueba de rachas para variables binomiales	893
Pruebas de rachas para datos cuantitativos	893
Coeficiente de correlación serial	893
Datos categóricos para dos muestras independientes.	
Tablas de contingencia	893
Diseños muestrales en tablas de contingencia de dos o más vías	894
Muestreo aleatorio simple o Modelo I	894
Muestreo aleatorio simple estratificado o Modelo II	895
Experimento aleatorio o Modelo III	896
Análisis de las tablas de contingencia de dos o más vías	896
Pruebas de Asociación o independencia	896
Modelo Lineal y Loglineal	896
Análisis Factorial de Correspondencias	897
Pruebas de independencia o asociación para tablas de contingencia	897
Grados de libertad para las tablas de contingencia	899
Corrección por continuidad	900
Otras medidas de asociación	900
Ejemplo	901
C20-1.SAS	904
Ejemplo	905
C20-2.SAS	908
Riesgo relativo y su intervalo de confianza	909
Ejemplo	911
Índice de desviación y su intervalo de confianza	912
Ejemplo	913

C20-3.SAS	914
Intervalos de confianza para las casillas	914
Ejemplo	915
Modelos lineales	916
Ejemplo	918
C20-4.SAS	918
C20-5.SAS	919
Análisis de Correspondencias	921
Ejemplo	922
C20-6.SAS	922
Partición de los grados de libertad en tablas de contingencia	925
Ejemplo	926
C20-7.SAS	926
Pruebas de independencia o asociación en los muestreos aleatorios simples	927
Modelos loglineales	927
Ejemplo	930
C20-8.SAS	930
Ejemplo	933
C20-9.SAS	934
Prueba exacta de Fisher para experimentos aleatorios o Modelo III.	939
Ejemplo	940
C20-10.SAS	941
Ejemplo	942
C20-11.SAS	943
Diferentes disposiciones de los datos en tablas de 2x2	944
Ejemplo	944
Tablas Multifactoriales y Análisis Múltiple de Correspondencias	945
Tablas Multifactoriales y Análisis Múltiple de Correspondencias en un Modelo Aleatorio simples estratificado o Modelo II (modelo lineal)	945
Ejemplo	947
C20-12.SAS	950
Tablas Multifactoriales y Análisis Múltiple de Correspondencias en un Modelo Aleatorio simples o Modelo I (Modelo log-lineal)	954
Ejemplo	954
C20-13.SAS	954
Análisis factorial de Correspondencias para el estudio de cualquier matriz de números positivos	961
Ejemplo	961
C20-14.SAS	962
Prueba $U$ de Mann-Whitney o prueba $S$ de suma de ordenaciones de Wilcoxon	963
Corrección por continuidad	964

Ejemplo	965
C20-15.SAS	967
Dos muestras apareadas o bloques aleatorios	968
Prueba de McNemar para significación de cambios	969
Ejemplo	970
C20-16.SAS	971
Prueba de los rangos con signos de Wilcoxon	972
Ejemplo	973
C20-17.SAS	974
Pruebas de hipótesis no paramétricas para más de dos muestras	975
Prueba de <i>Kruskal-Wallis</i> para más de dos muestras independientes.	975
Ejemplo	976
C20-18.SAS	977
Prueba de <i>Bartlett</i> para homogeneidad de varianzas	978
Bloques aleatorios o muestras relacionadas para más de dos niveles del tratamiento	978
Prueba <i>Q</i> de <i>Cochran</i> para <i>t</i> muestras relacionadas	978
Ejemplo	979
C20-19.SAS	980
Prueba de <i>Friedman</i> para diseños de bloques aleatorios	980
Ejemplo	980
C20-20.SAS	981
Pruebas no paramétricas con dos variables. Correlación no paramétrica	982
Coeficiente de correlación de rangos de <i>Spearman</i>	982
Coeficiente de correlación de clasificación de <i>Kendall</i>	982
Coeficiente de correlación serial	982

<b>TABLAS</b>	985
---------------	-----

<b>Índice Analítico</b>	1019
-------------------------	------

# **CAPÍTULO 1**

## **Introducción**



# CAPÍTULO 1

## Introducción

La justificación del título del presente manual es bien simple, este manual está orientado en la descripción, manejo y análisis de resultados de las aplicaciones (eminentemente prácticas, remacharíamos con redundancia) de los métodos estadísticos, sin detenemos en el origen matemático de dichos métodos que, ni por formación ni por necesidad, se consideran en el presente manual.

### **Estadística.-**

La concepción más difundida de la Estadística es la de su primer origen, es decir, la estadística como la actividad de recuento y enumeración que tradicionalmente ha sido realizada por el Estado con diversas finalidades. Estas primeras técnicas estadísticas consistían esencialmente en la organización de datos, su representación gráfica y en el cálculo de cantidades representativas del conjunto, con el objeto de que los aspectos sobresalientes del mismo fueran rápida y fácilmente comprensibles. Esta parte de la Estadística es lo que, en la terminología moderna, se conoce como *Estadística Descriptiva*. Y si bien éste es un aspecto interesante de la Estadística, no es el único y en la concepción moderna de esta disciplina tampoco es el más importante.

La Estadística Descriptiva ha visto reducida su importancia porque sus técnicas son de utilidad limitada cuando se tiene grandes poblaciones o poblaciones numéricamente infinitas o poblaciones finitas que se podrían estudiar en toda su amplitud pero sería enormemente costoso para los fines propuestos, por lo que se estudia sólo una parte representativa de la población, denominada *muestra*, y las conclusiones se generalizan a la *población*. Este aspecto de la Estadística es el que se conoce como *Estadística Inferencial* o *Estadística Inductiva* y por ser, hoy día, el aspecto más importante de esta ciencia es a lo que se refiere el término **ESTADÍSTICA**, sin calificativos.



## Estadística y cuantificación de la incertidumbre.-

En el epígrafe anterior se ha realizado una primera aproximación a la distinción entre *muestra* y *población*, y se ha hecho notar que el interés de la Estadística, como método de investigación, se centra en las inferencias, las cuales van más allá de las observaciones que forman una muestra dada. Por tanto, cualquier conclusión a la que se llegue contendrá elementos de *incertidumbre*. Esta incertidumbre es inherente a cualquier proceso en el que se extrapolan conclusiones de un subconjunto al conjunto.

Expliquemos con ejemplos estos tres conceptos (población, muestra e incertidumbre). En todas las ciencias empíricas se plantean problemas que afectan a la población completa. Pongamos como ejemplos y para una mejor comprensión, seis problemas muy simples y conocidos.

- 1) ¿Qué relación existe entre las longitudes de los lados de un triángulo dos de cuyos lados forman un ángulo recto?
- 2) ¿Cuánto tiempo tarda en llegar al suelo un objeto dejado caer desde un cuarto piso ( $12m$  de  $h$ )?
- 3) ¿Cuántos movimientos respiratorios por minuto realiza el caballo?
- 4) ¿Cuál es la concentración de glóbulos rojos en la sangre humana?
- 6) ¿Cuánto tiene que pesar un niño de dos años y qué cantidad de agua diaria necesita?
- 7) ¿Cuál la altura de una persona a los 18 años?

Esta claro que estos son problemas que se plantean en la población, es decir, no se quiere saber cuál es la relación entre los lados de un triángulo rectángulo concreto, ni de dos, se quiere tener una respuesta para todos los triángulos rectángulos habidos y por haber. Lo mismo ocurre para el tiempo de caída, se quiere una respuesta para todos los objetos habidos y por haber y sea cual sea el tiempo y lugar en que ocurra dicha caída. Y lo mismo ocurre con los movimientos respiratorios del caballo, la concentración de glóbulos rojos, el peso de niños de dos años o la altura de personas de 18 años la respuesta tiene que ser para todos los caballos, todos los seres humanos, todos los niños de dos años o todas las personas de 18 años, habidos y por haber, no para los veinte caballos con los que se ha hecho la experiencia o las 100 personas a las que se le ha medido la concentración de glóbulos rojos o los 5000 niños y personas de 18 años que se han pesado y medido.

Pero las cuestiones anteriores pueden plantearse también en forma de hipótesis de la siguiente manera

- 1) ¿En un triángulo rectángulo, la suma de los cuadrados de los lados menores es igual al cuadrado del lado mayor?
- 2) ¿Un objeto dejado caer desde un cuarto piso ( $12m$  de  $h$ ), tarda  $1.565$  seg en llegar al suelo?
- 3) ¿Realiza 12 movimientos respiratorios por minuto el caballo?
- 4) ¿Es de  $5.200.000/mm^3$  la concentración de glóbulos rojos en la sangre humana?
- 6) ¿Pesa  $11.8$  Kg un niño de dos años?
- 7) ¿Mide  $1.75$  m de altura una persona a los 18 años?

¿Porqué nos planteamos estas hipótesis como respuesta a los problemas anteriormente surgidos? No sería suficiente con pesar un niño de dos años o medir una persona de 18 años o contar los glóbulos rojos de una persona o medir los lados de un triángulo rectángulo, o medir el tiempo que tarda en caer un objeto desde el cuarto piso, *etc.* Claramente la respuesta es que no. ¿Porqué?

En la naturaleza hay variabilidad

Esto quiere decir que una medida realizada en diferentes experiencias da valores diferentes por tres motivos diferentes.

1) Si se mide una misma persona de 18 años dos veces diferentes (por dos personas o por la misma persona pero haciéndole ver que se esta midiendo a dos personas diferentes) probablemente dará medidas diferentes; si se pesa un niños de dos años dos veces diferentes ocurrirá lo mismo; si se miden los lados de una parcela con forma de triángulo rectángulo, ocurrirá lo mismo; y lo mismo se puede decir de las medidas de las otras variables definidas anteriormente, es decir, del número de movimientos respiratorios en el caballo, de la concentración de glóbulo rojos en la sangre y del tiempo de caída de un objeto desde el cuarto piso.

Esta falta de coincidencia entre el autentico valor de una observación y el valor realmente medido, es lo que se denomina

Error de medida

Este error de medida va implícito, y es inseparable, a toda variable cuantitativa, y es debido a los diferentes estados de animo de la persona que realiza la medida, a su agudeza visual, a su velocidad de reflejos, a su nivel de cansancio, a su habilidad para manejar los instrumentos de medida, *etc.*

2) Si se miden diez personas de 18 años probablemente dará medidas diferentes; si se pesan diez niños de dos años ocurrirá lo mismo; si se miden los lados de diez parcelas con forma de triángulo rectángulo, ocurrirá lo mismo; y lo mismo se puede decir de las medidas de las otra variables definidas anteriormente, es decir, del número de movimientos respiratorios en diez caballos, de la concentración de glóbulo rojos en la sangre de diez personas y del tiempo de caída de diez objetos desde el cuarto piso.

Esto es así porque, además del error de medida, en la altura de las personas de 18 años intervienen muchas causas: fondo genético (altura de sus padres), régimen alimenticio seguido, altitud en la que se ha criado, enfermedades padecidas,... Estas diez personas medirían lo mismo si las diez hubieran tenido las mismas experiencia vitales desde la concepción a la actualidad, es decir, si las diez fueran la misma persona. Incluso los gemelos univitelinos, que tienen el mismo fondo genético, tendrán diferentes alturas porque todas las demás causas serán diferentes en ambos. Lo mismo se puede decir de las otras variables.

Esta falta de coincidencia entre el valor que tendría que tener una variable en una población y el valor realmente medido en diferentes unidades, es lo que se denomina

Error de muestreo

Este error de muestreo es debido a la falta de uniformidad en las causas que intervienen en la expresión del valor de una variable. Este error refleja el hecho de la *individualidad biológica*, es decir, refleja el hecho de que diferentes individuos miden diferente aunque no existiera el error de medida.

3) Cuando se midan las personas de 18 años o se pesen los niños de dos años, *etc.*, medirán diferente no solo por el error de medida y el error de muestreo, sino también por el

Error experimental

Este error experimental es debido a la falta de uniformidad en las condiciones de las experiencias. Por ejemplo, en la altura de personas de 18 años puede ser debido a que las personas que se están midiendo no tienen exactamente 18 años, sino que hay meses de diferencia entre las edades de los individuos lo que hace que mayores midan más que lo más jóvenes.

Si existe esta variabilidad en la naturaleza como se puede llegar a saber el valor *esperado* de una variable que nos permita realizar predicciones fiables, es decir, que sea un valor poblacional correcto.

Hay dos soluciones:

- 1) La primera es que se pueda llegar a dicho valor analíticamente, es decir, *deductivamente*, con lo cual el resultado es incuestionable.
- 2) La segunda, para el caso que no se puede deducir su valor poblacional, es estudiar toda la población o estudiarla *inductivamente* (estadísticamente).

Explicemos estos dos términos, el de las deducciones y el de las inferencias inductivas o razonamiento inductivo.

El proceso lógico por el que se llega a generalizaciones partiendo del estudio de casos particulares se denomina *inducción*, opuesto al de la *deducción*, que implica la extracción de conclusiones particulares o especializadas partiendo de proposiciones generales.

***Razonamiento deductivo es el que va de lo general a lo particular.***

***Si*** los seres vivos están compuestos de células.

***Y*** todos los gatos son seres vivos.

***Entonces*** todos los gatos están compuestos de células.

La conclusión de este razonamiento está implícita en las dos afirmaciones anteriores. Si las premisas son ciertas, la conclusión no puede discutirse. En realidad, la conclusión no añade nada, aunque la cadena del razonamiento puede resultar muy valiosa al revelar verdades que están implícitas en las premisas.

Con razonamientos como este se puede llegar a demostrar que el cuadrado del lado mayor de un triángulo rectángulo es igual a la suma de los cuadrados de los lados menores, sin necesidad de medir los lados de ningún triángulo rectángulo, por lo que la primera de las hipótesis anteriormente planteadas ya está resuelta, y este resultado es indiscutible. Lo mismo ocurre con la segunda hipótesis, con razonamientos deductivos se llega a la conclusión de que el tiempo que tarda en llegar al suelo un objeto dejado caer desde una altura  $h$  es igual a la raíz cuadrada de dos veces esta altura dividida por el valor de la gravedad en la zona que se esté haciendo la experiencia. Por lo que se puede saber dicho tiempo sin necesidad de realizar la experiencia, y el resultado de esta deducción es incuestionable.

Una vez deducida la relación entre los lados de un triángulo rectángulo, cuando se plantea un problema de medición de tierra, por ejemplo, en el que esté implicado un triángulo de este tipo, se esperará un valor para el lado mayor medidos los lados menores aunque una vez realizadas las medidas no coincidan exactamente como consecuencia de los errores anteriormente citados. Lo mismo se puede decir del tiempo de caída de un objeto desde un cuarto piso, el tiempo esperado es de 1.565 seg, aunque si se realiza la experiencia la medida realizada, probablemente difiera de este valor por los errores anteriormente citados. En todo caso, las medidas reales estarán alrededor del valor esperado.

En las otras cuatro experiencias no pueden demostrarse deductivamente los valores esperados ¿Cuál es el número de movimientos respiratorios del caballo en condiciones de normalidad? No hay una fórmula que nos permita deducir el valor de esta variable. Por lo que habría que estudiar todos los caballos habidos y por haber sobre la faz de la tierra para dar respuesta a esta cuestión. O bien estudiar solo unos pocos caballos (muestra) y a partir de ellos inferir los valores a la totalidad de los caballos (población).

### ***Razonamiento inductivo es el que va de lo particular a lo general.***

***Si*** el animal  $A$ , el animal  $B$ ,... están compuestos de células.

***Y*** el animal  $A$ , el animal  $B$ ,... son gatos.

***Entonces*** todos los gatos están compuestos de células.

Como se observa, la conclusión es más amplia que las premisas. Se añade algo nuevo. Si la conclusión es cierta, el conocimiento humano ha sido ampliado. Pero no existe certeza de que las conclusiones del razonamiento inductivo sean verdaderas, estas conclusiones valen solamente en términos de probabilidad y deben ser verificadas o contrastadas con los resultados de la observación; si se mantienen deberá ser sobre la base de su compatibilidad con los hechos de la naturaleza en el ámbito del conocimiento en el que se desenvuelve el razonamiento.

El razonamiento inductivo es inseparable de cualquier rama del conocimiento, científico o no. El conocimiento común, esto es, nuestras opiniones, actitudes y, por supuesto, nuestros prejuicios están basados en informaciones incompletas y razonamientos de tipo inductivo. Pongamos otro ejemplo inductivo semejante al anterior: Si el animal A, el animal B, ..., son animales ariscos, y el animal A, el animal B, ... son gatos; entonces todos los gatos son ariscos. Otro ejemplo: si la letra con la que esta escrita la receta del médico de cabecera no la entiendo, la receta del cardiólogo no la entiendo, la receta del reumatólogo no la entiendo, ... , y el médico de cabecera, el cardiólogo, el reumatólogo, ... , son médicos; entonces los médicos tienen una escritura ilegible.

Está claro que se puede llegar a estos tipos de conclusiones porque todos los gatos que se hayan conocido hasta el momento sean ariscos y todos los médicos que nos hayan atendido tengan una letra difícilmente legible. Pero también es claro que no es imposible, si se sigue observando, el encontrar gatos cariñosos y médicos que tengan una letra perfectamente legible.

Esta es la incertidumbre típica de los procesos inductivos, y esta incertidumbre existe, como se ha dicho anteriormente, porque existe variabilidad en la naturaleza. Si todos los caballos hicieran el mismo número de movimientos respiratorios por minuto, o todos los individuos de 18 años midieran lo mismo, o todos los niños de dos años pesaran lo mismo o todos los individuos tuvieran exactamente la misma concentración de glóbulos rojos en la sangre, no habría esta incertidumbre.

La diferencia entre el conocimiento científico y el conocimiento común es que mientras el conocimiento común no cuantifica el grado de incertidumbre, el conocimiento científico si cuantifica la incertidumbre, y es la Estadística la encargada de esta cuantificación.

La cuantificación de la incertidumbre se logra mediante el uso de los conceptos y técnicas de la Probabilidad.

### **La Estadística y el Método Científico.-**

En los epígrafes precedentes se ha esbozado la función de la Estadística en la presentación y organización de datos y en la cuantificación de la incertidumbre en el proceso inductivo cuando esta forma parte del conocimiento científico. Pero también es esencial la participación de la estadística en el proceso *hipotético/deductivo*.

Esto quiere decir que la estadística sirve también para averiguar la realidad de pequeñas, pero significativas, diferencias en algunos de los factores causantes de la variabilidad, lo que nos permitirá, de esa manera, detectar cuales son dichos factores y, una vez conocidos, saber el grado de importancia que tienen diferentes factores en la determinación de un carácter medido como variable.

La consecuencia de la incertidumbre en el proceso inductivo es que los fenómenos naturales, que en teoría son fundamentalmente determinísticos, se nos presenten como probabilísticos como consecuencia de la variedad y multitud de

factores causales y nuestra torpeza para controlar éstos factores. Por medio de la estadística se puede cuantificar esa probabilidad y sacar a la luz la manifestación de fenómenos que a simple vista parecen aleatorios.

Pongamos un ejemplo muy simple de porqué es necesaria la utilización de la estadística en la cuantificación de la incertidumbre del proceso inductivo.

Si se quiere saber si esta bien hecho un dado se puede tomar una muestra, de 600 tiradas, de las infinitas tiradas que constituye la población de dicho dado. Si en estas 600 tiradas salen 600 *unos* concluimos, sin más, que el dado está mal hecho. Si, por el contrario, salen 100 *unos* concluimos, sin más, que el dado esta bien hecho. Pero la naturaleza, como consecuencia de los errores anteriormente expuestos, nunca se manifiesta con esa claridad. Si lanzamos 600 veces un dado, muy probablemente nunca salga ni 100 ni 600 *unos*, sino que saldrá un número intermedio, por ejemplo 120 *unos*. El que salga 120 *unos* se nos manifiesta como algo aleatorio pero puede ser la manifestación del 1/6 esperado. La estadística nos ayuda a tomar una decisión sobre si es o no el 1/6 esperado, y por lo tanto, nos ayuda a decidir sobre si está o no está bien hecho el dado.

Por lo tanto, cualquier caracterización que se haga con los datos disponibles involucra necesariamente un razonamiento de tipo deductivo, mientras que la generalización de las conclusiones se hace por razonamiento inductivo, de manera que los dos tipos de razonamientos lógicos aparecen mezclados en la metodología estadística, como consecuencia de que la metodología estadística no es sino la metodología de todas las ciencias empíricas, es decir, lo que se conoce como *Método Científico (Método hipotético-deductivo)*. Lo que la diferencia a una ciencia de otra ciencia es el diferente aspecto del *objeto científico* que se estudia, y la estadística se utilizará, en las diferentes ciencias, como una técnica o herramienta para conocer dicho objeto.

Y pongamos otro ejemplo para aclarar la utilización de la estadística en la detección de pequeñas, pero significativas diferencias en algunos de los factores causantes de la variabilidad

Supóngase que un experto en nutrición considera que dada la composición de cierto nuevo producto alimenticio, presenta ventajas nutricionales sobre el producto estándar (hipótesis 1). Por lo tanto el experto deduce ventajas que ha de contrastar con la realidad. Para ello diseña un ensayo de nutrición con, por ejemplo, diez animales alimentados con el nuevo producto y diez con el estándar. No obtiene el resultado esperado (inducción 1), y repasando las condiciones de aplicación de los ensayos deduce que la concentración del nuevo producto ha sido excesiva (hipótesis 2). Para ello diseña un ensayo semejante al anterior pero con una concentración inferior del nuevo producto, obteniendo un resultado algo mejor (inducción 2). Deduce que el resultado no es óptimo porque no se ha suministrado la cantidad adecuada de agua (hipótesis 3), por lo que diseña un tercer ensayo del que obtendrá la inducción 3,...

Por lo que se puede definir, nuevamente, la Estadística como *el estudio científico de fenómenos naturales basado en datos numéricos que se apartan de las leyes de la causalidad exacta.*

Desmembremos esta definición:

- Estudio científico es el estudio sistemático de un fenómeno con el fin de descubrir nuevos hechos o principios que permitan interpretar y transformar la realidad que nos rodea. Requiere los siguientes elementos:

- 1) Hechos observados, con los que se establece un problema (un nuevo producto alimenticio con una composición diferente a la estándar).
- 2) Hipótesis o idea provisional de cómo los hechos han de ser interpretados y explicados (concentración del nuevo producto o cantidad de agua que ha de acompañarle en la dieta). Siempre es necesaria una hipótesis previa para la obtención de datos que sean relevantes al problema, aunque en algunas ocasiones un estudio previo de Estadística Descriptiva pueden sugerir hipótesis apropiadas.
- 3) Obtención de datos por medio de experimentos, ensayos u observaciones dirigidas, destinados a probar la validez de la hipótesis (diez animales criados con una dieta y otros diez criados con otra dieta, siendo todo los demás factores los mismos para los dos grupos). En este aspecto del método, el papel de la Estadística es esencial. Se trata de adquirir información de manera que ésta sea relevante al problema y las conclusiones que se extraigan tengan cierto grado de fiabilidad. La cantidad de información necesaria, la forma de recolectarla y las técnicas para adquirirla, de manera que se cumplan los dos objetivos de información relevante y fiabilidad, son todos problemas que resuelve el método estadístico. La rama de la Estadística que trata este problema es la de los *Diseños Experimentales*.
- 4) Resultados y su interpretación, con los que se rechaza o no la hipótesis postulada. Una vez obtenidos los datos, el papel de la Estadística se hace aún más importante, puesto que estos datos hay que analizarlos. El primer paso, en este análisis, puede ser la organización de los mismos, su presentación gráfica y su descripción, resaltando los aspectos más característicos (*Estadística Descriptiva*). La segunda y más importante parte del análisis es la de la generalización de los aspectos característicos de la información y examinar su compatibilidad con las consecuencias de la hipótesis. Aunque la información será incompleta, tratará de reflejar la variabilidad del fenómeno estudiado, realizando la generalización acompañada de un grado de incertidumbre cuantificable probabilísticamente (*Estadística Inferencial*).

- Fenómenos naturales son todos aquellos fenómenos que ocurren en la naturaleza, animada o inanimada, fuera del control del hombre, más aquellos provocados por el científico (experimento) y que en parte están bajo su control.

- Datos: los estudios estadísticos se realizan grupos de individuos y no con un solo dato.

- Numéricos: si los datos no pueden cuantificarse de una manera u otra no tendrán utilidad para el análisis estadístico.

- Causalidad exacta: como se ha dicho, la mayor parte de los fenómenos naturales son influidos por muchos factores causales, incontrolables en su variación y muy a menudo no identificables. La Estadística permite medir tales factores variables con un error predecible.

Y averiguar la realidad de pequeñas, pero significativas, diferencias. Los fenómenos naturales son fundamentalmente determinísticos y únicamente la variedad de factores causales y nuestra torpeza para controlar éstos factores hacen que dichos fenómenos aparezcan como probabilísticos.

### **Bibliografía:**

- Infante Gil, S. y Zárate De Lara, G.P.* 1984. MÉTODOS ESTADÍSTICOS. Ed. TRILLAS. México.
- Lite, TM, y Jackson Hills, F.* 1987. MÉTODOS ESTADÍSTICOS PARA LA INVESTIGACIÓN EN LA AGRICULTURA. Ed TRILLAS. México.
- Sokal, R.R. y Rohlf, F.J.* 1994. BIOMETRY. Ed. W.H.FREEMAN. San Francisco.
- Snedecor, G.W. y Cochran, W.G.* 1971. MÉTODOS ESTADÍSTICOS. Ed C.E.C.S.A. México.
- Steel, R.* 1996. PRINCIPLES AND PROCEDURES OF STATISTICS. Ed. McGRAW-HILL Education. New York .





# **CAPÍTULO 2**

## **Fundamentos**



# CAPÍTULO 2

## Fundamentos

### Distribuciones.-

La forma más elemental de presentación de una variable aleatoria podría consistir en una simple enumeración de todos los posibles valores ordenados en sentido creciente. Es lo que se denomina una *serie estadística*. Esto será posible muy pocas veces pues, habitualmente, se tratará con poblaciones infinitas, en las que los posibles valores de la variable serán infinitos o serán finitos pero se repetirán un número de veces.

En otra ocasiones ocurrirá que la variable tomará ciertos valores que se repiten, en cuyo caso se puede realizar una *distribución*, es decir, anotar cada valor que toma la variable una sola vez y al lado la probabilidad con la que se presenta dicho valor (*frecuencia* si es una muestra).

Pero si el número de valores que puede tomar la variable es muy elevado (infinito), para la representación o caracterización de la distribución, se impone el agrupar los posibles valores de la variable en *clases* o categorías (por muy pequeñas que estas sean), caracterizadas por los *límites de clase*, por la *marca de clase* y por la *densidad de probabilidad* (frecuencia o número de individuos) de cada clase, esto es, una *función de probabilidad*.

### Función de probabilidad.-

Toda variable,  $X$ , tiene una *función de densidad de la probabilidad* o *función de probabilidad* o *función de densidad*,  $f(x)$ , que proporciona la probabilidad que tiene dicha variable aleatoria de tomar un valor determinado del *campo de variación*, *rango* o *amplitud*.

Adviértase que esto es análogo a la distribución de frecuencias relativas en la que las frecuencias relativas han sido sustituidas por probabilidades. Así se puede pensar que las *distribuciones de probabilidad* son teóricamente las *distribuciones de*

*frecuencias relativas* llevadas al límite, es decir, cuando el número de observaciones hechas es muy grande. Por esta razón se pueden ver las *distribuciones de probabilidad* como *distribuciones de poblaciones*, y las *distribuciones de frecuencias relativas* son *distribuciones de muestras*.

La expresión *variable aleatoria* se usa para variables que poseen funciones de densidad de probabilidad.

### **Función de distribución.-**

La función de distribución,  $F(x)$  es la probabilidad que tiene una variable aleatoria de tomar un valor igual o menor que uno determinado del campo de variación. Es decir, es la función asociada a la distribución de probabilidad acumulada.

La función de distribución varía entre 0 y 1, puesto que la probabilidad de que el valor de la variable sea menor o igual que el último valor del rango, es la suma de todas las frecuencias relativas, es decir, es 1.

### **Experimento, espacio muestral y suceso.-**

En ciencia, los datos son colecciones de observaciones obtenidas estudiando el comportamiento de un fenómeno, ya sea en estado natural o bien bajo control humano.

Un experimento es el proceso mediante el cual se obtiene una observación del carácter o medida de la variable.

Y un experimento aleatorio es aquél cuyos resultados no pueden predecirse antes de su realización (principio de incertidumbre) por estar sujeto a múltiples causas o errores, manifestándose como aleatorio.

El conjunto formado por todos los resultados posibles de un experimento o posibles valores de la distribución, recibe el nombre de *espacio muestral*, que se identificará con la letra  $M$ .

### **Ejemplo.-**

Se mide el volumen relativo de glóbulos rojos con respecto al volumen total de sangre. Cualquiera de los infinitos valores (dependiendo de la precisión de la medida) entre cero y cien es un posible resultado del experimento. El espacio muestral es

$$M = (X_i = 0 \leq X \leq 100)$$

## Ejemplo.-

Se mide la reacción alérgica a ciertos producto ambiental mediante inyección subcutánea. Se registra la respuesta como favorable ( $f$ ) o como desfavorable ( $d$ ). El espacio muestral es

$$M = (f, d)$$

Como se observa en estos dos ejemplos, el espacio muestral puede ser *continuo* o *discreto*

y no necesariamente tiene que estar formado por números; puede estar formado por *atributos* o *cualidades*.

Un *evento* o *suceso* es un subconjunto de un espacio muestral.

Si un espacio muestral es discreto, se tienen tantos sucesos como valores pueda tomar la variable. En el ejemplo del alergeno, los sucesos son o  $f$  o  $d$ .

En el caso del espacio muestral continuo los sucesos son infinitos, por lo que será conveniente dividir el espacio muestral en cierto número de clases que sean exhaustivas y mutuamente excluyentes, lo que constituirán los sucesos de este espacio muestral infinito.

En el experimento del volumen relativo de glóbulos rojos, cualquiera de los infinitos valores entre cero y cien es un suceso, por lo que se puede hacer una partición de este espacio muestral en, por ejemplo, tres sucesos, tal que

$$A = (0 \leq X < 25.5)$$

$$B = (25.5 \leq X < 50.5)$$

$$C = (50.5 \leq X < 75.5)$$

$$D = (75.5 \leq X \leq 100)$$

Según el tipo de espacio muestral que se tenga, se tendrá un tipo de función y un tipo de variable. Estos son *discretas* y *continuas*.

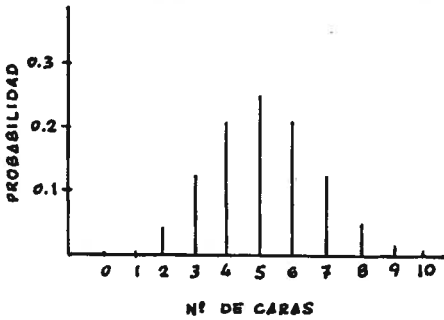
El tipo de espacio muestral indicará los dos tipos fundamentales de distribuciones.

## Distribuciones de tipo discreto.-

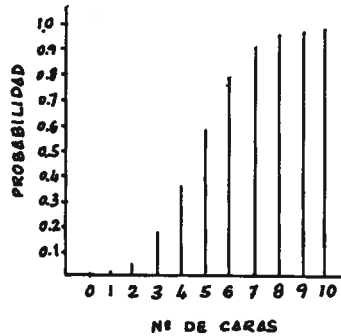
Es aquella en la que el espacio muestral de la variable aleatoria solo puede tomar ciertos valores, que normalmente serán números enteros o atributos, que se asociaran a todos los posibles sucesos.

Se representa siempre con un *diagrama de barras*, pues los posibles valores están perfectamente diferenciados y aislados los unos de los otros, y el tamaño de la barra es la probabilidad de dicho suceso. De manera que la función de probabilidad y la función de distribución del número de caras que pueden salir al lanzar una moneda

diez veces es



función de probabilidad



función de distribución

Como se ve, la función de densidad de una variable discreta  $X$  es el conjunto de pares ordenados

$$X_i, f(X_i) = P(X_i)$$

donde  $X_i$  es cada uno de los valores que puede tomar la variable o cada uno de los sucesos, y  $f(X_i)$  es la probabilidad asociada con el valor particular  $X_i$ .

Puesto que en el caso de una variable discreta los posibles valores de  $X$  tienen correspondencia con elementos concretos (sucesos) del espacio muestral, es claro que deben cumplirse las siguientes propiedades

$$f(X_i) \geq 0 \text{ para todo valor } X_i$$

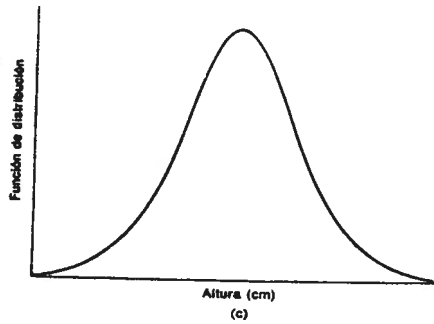
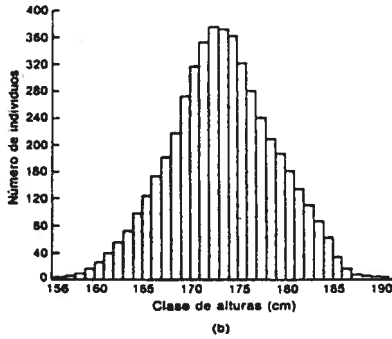
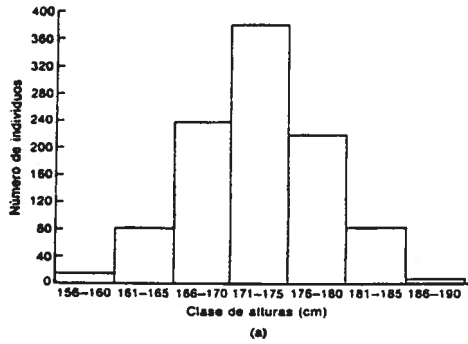
$$\sum_i f(X_i) = 1 \text{ donde la suma es para todo valor } X_i \text{ que puede tomar la variable } X.$$

### Distribuciones continuas.-

Sin embargo para los caracteres cuantitativos se requiere un modo distinto de descripción.

El concepto básico es el de **distribución estadística**. Si se miden las tallas de un gran número de estudiantes varones con una precisión de 5 cm, diferirán unas de otras, entre 155 y 190, pero caerán muchos más individuos en las categorías intermedias, entre 170 y 175, que en las extremas.

Representando cada clase observada por una barra, de altura proporcional al número de individuos de la clase, el resultado es una gráfica como la (a).



En esta gráfica, el número de individuos contados en función de la clase a la que pertenecen es un **histograma de frecuencias**.

Supóngase ahora que se miden cinco veces más individuos, cada uno al centímetro más próximo. Las clases de la gráfica (a) se tienen que subdividirse para producir un histograma como el b.

Si se continua este proceso, aumentando la precisión de la medida, pero incrementando proporcionalmente el número de individuos medidos, el histograma adoptará eventualmente la apariencia continua de la gráfica c que es la **función de distribución** de las alturas en la población.



Esta curva continua es una idealización, ya que no se puede tomar ninguna medida con una precisión infinita, ni se puede aumentar infinitamente el número de individuos medidos.

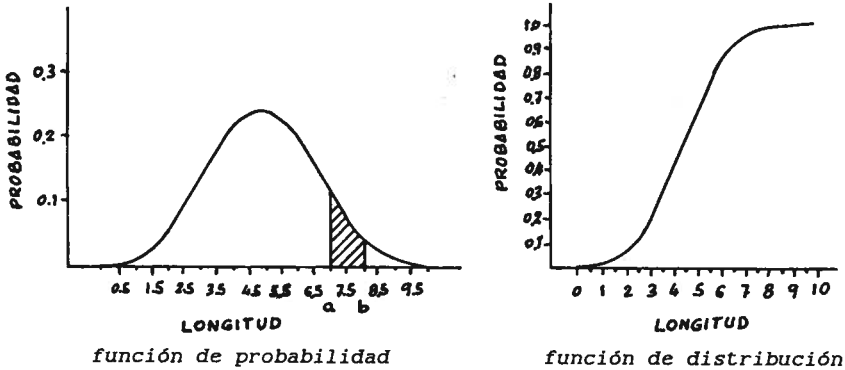
Además la propia variable que se va a medir puede ser intrínsecamente discontinua porque consista en enumerar objetos discretos, como las facetas del ojo o las quetas.

Sin embargo, es conveniente desarrollar el análisis mediante este cuadro ligeramente idealizado, como si fuera una notación simplificada del histograma de frecuencias reales (**b**) que es más complicado.

Por lo tanto, las distribuciones continuas son aquellas en las que la variable aleatoria puede tomar todos los valores posibles dentro del campo de variabilidad o rango. Es decir, son aquellas en las que el campo muestral es infinito.

Si se pretende representar, la primera dificultad es que no se puede representar, como ocurría con la anterior, con una barra para cada posible sucesos, pues se necesitarían infinitas barras.

En lugar de barras se necesitaría un *histograma*, esto es, el dividir el campo muestral en clases mutuamente excluyentes que constituirá un rectángulo del histograma tan alto como la probabilidad de que ocurra el suceso definido por dicha clase. Si estas clases tienden a infinito, la representación sería



De modo que la única forma que se tiene de caracterizar la distribución de probabilidades es mediante una ecuación y representarla gráficamente como una forma límite del histograma de frecuencias relativas.

La función de densidad de tipo continuo debe cumplir las siguientes propiedades

$f(X_i) \geq 0$ , es decir, que la curva siempre se encuentra sobre el eje de abscisas.

El área total bajo  $f(X_i)$  es 1.

El área delimitada por dos líneas verticales levantadas sobre los puntos  $a$  y  $b$  ( $a < b$ ) y la función, es la probabilidad de que  $X$  tome un valor entre  $a$  y  $b$ .

Una diferencia fundamental entre funciones de densidad de tipo discreto y continuo es que, en el caso de las discretas, la probabilidad total se encuentra distribuida en un número finito de puntos concretos, conocidos.

Como consecuencia, si la variable puede tomar un valor  $k$ , puede asignarse un valor a  $f(X_i)=P(k)$ , con la condición de que la suma de probabilidades para todos los puntos sea uno.

Si se trata de una variable continua, ésta puede tomar infinitos valores en un intervalo dado, y si se le asignara una probabilidad, por muy pequeña que sea, a cada uno de estos infinitos valores, la probabilidad total sumaría más de uno.

Esto lleva a concluir que, para una variable aleatoria continua, la probabilidad asociada con un punto cualquiera es cero. Es decir,  $P(k)=0$  para cualquier valor que tome la variable, consecuentemente:  $f(X_i) \neq P(X_i)$ .

A simple vista la afirmación anterior puede parecer desconcertante, puesto que podría interpretarse, por ejemplo, como que ninguna persona pueda tener una estatura de 1.62 m ya que  $P(X=1.62)=0$ , según se ha afirmado en el párrafo anterior.

Esta aseveración se debe interpretar como que, dada la precisión de nuestros instrumentos de medición, no se puede distinguir entre una persona que mida 1.62 o 1.624 o 1.6235 o 1.6245, pero no importa porque en lo que se esta realmente interesado es en la probabilidad del intervalo 1.615/1.625 (como ya se ha visto anteriormente).

Por otra parte, puesto que las probabilidades son áreas bajo la curva, sólo es posible calcular probabilidades (áreas) correspondientes a intervalos dados, ya que, por definición, el área de una línea (que correspondería a la probabilidad de un solo valor, por ejemplo, que una persona mida 1.62 m) es cero.

En otras palabras, cuando se mide una persona (o cualquier otra variable continua) y se da el valor de 1.62, no se está seguro de ese valor, pero si se tiene que estar seguro de que esa persona mide más de 1.615 y menos de 1.625, que son los límites de clase de esa medida.

Límites reales que delimitan el suceso y hace posible hallar la probabilidad o área de dicho suceso.

Por tanto, cuando se tome la medida de una variable continua el último decimal tiene que denotar precisión con objeto de poder determinar los límites de clase añadiendo un decimal más, y que estos sean mutuamente excluyentes.

A partir de ahora, toda variable continua que se estudie se le supondrá una

distribución normal, si no se demuestra lo contrario.

### **Distribuciones como modelos para representar situaciones reales.-**

Ya se ha hecho notar la relación entre las frecuencias relativas e histogramas y las funciones de densidad. Pudiéndose citar las primeras como *distribuciones empíricas* y las segundas como *distribuciones teóricas*. Por regla general, en la práctica se tendrá distribuciones empíricas que se asociaran a distribuciones teóricas que sirvan de modelo para entender el fenómeno que se está estudiando.

Una práctica corriente en la Ciencia es tratar de explicar fenómenos naturales mediante *modelos*. Un modelo es una representación reducida de la realidad, es decir, de la complejidad de la realidad se modeliza la esencia de esta, la que se quiere explicar. Dicho de otro modo, un modelo es una interpretación de un hecho abstracto o inobservable expresado en términos de hechos no abstractos u observables. Por ejemplo, se puede decir que la *Tierra es como un globo* (esférica) y gira sobre un eje Norte-Sur. Que la Tierra es redonda es un hecho conjeturable y por lo tanto inferible pero no es un hecho observable (nadie ha visto la totalidad del planeta, ni siquiera los astronauta pueden verla en su totalidad) y para comprender ciertos hechos se utiliza un modelo no abstracto y por lo tanto observable, esto es, que la Tierra es como el globo que tenemos sobre la mesa. Este es un modelo verbal que permite comprender y explicar multitud de observaciones y hechos empíricos, y permite prever resultados de experiencias aún no realizadas, como puede ser la de navegar siempre hacia el Oeste para volver al lugar de partida, lo que confirmará la veracidad (o falsedad) del modelo.

Modelos más perfeccionados dependen de afirmaciones matemáticas que permiten explicar los hechos observables (los datos), por medio de hechos no observables (parámetros, *efecto de un factor*, etc.). El aspecto más importante de un modelo matemático es que posee las características intrínsecas del problema real.

Se puede pensar que los fenómenos naturales son fundamentalmente determinísticos y únicamente la variedad y multitud de factores causales y nuestra torpeza para controlar éstos factores hacen que dichos fenómenos se nos manifiesten como probabilísticos. Por medio de la estadística se puede cuantificar esa probabilidad y sacar a la luz la manifestación de fenómenos que a simple vista parecen aleatorios.

Por lo que las razones para utilizar modelos probabilísticos se derivan de la idea de que en los experimentos científicos el resultado exacto no puede preverse de antemano, porque la mayor parte de los fenómenos naturales son influidos por muchos factores causales, incontrolables en su variación y muy a menudo no identificables. Sin embargo, el fenómeno que se está estudiando en los experimentos, se hará a través de la medida de una variable de la que sí se prevé su espacio muestral con los posibles resultados del experimento aleatorio como consecuencia de la función de probabilidad que describe el comportamiento de esta variable.

Pongamos un ejemplo muy simple, si se tira un dado al aire no se sabe cual número va a salir pero sí se sabe cual es el espacio muestral y que saldrá un suceso de este espacio con una probabilidad de  $1/6$ . Si se está realizando experimentos con dados, se asume dicha distribución (binomial de  $p=1/6$ ; ver más adelante) pero dicha

probabilidad no se nos manifiesta claramente, hay que asumirla o hay que probarla de la siguiente manera: se toma una muestra, por ejemplo, de 600 tiradas, de las infinitas tiradas que constituye la población de dicho dado. Si en estas 600 tiradas salen 600 unos concluimos, sin más, que el dado está mal hecho ( $p \neq 1/6$ ). Si, por el contrario, salen 100 unos concluimos, sin más, que el dado está bien hecho ( $p = 1/6$ ). Pero la naturaleza, como consecuencia de los errores anteriormente expuestos, nunca se manifiesta con esa claridad. Si lanzamos 600 veces un dado, muy probablemente nunca salga ni 100 ni 600 unos, sino que saldrá un número intermedio, por ejemplo 120 unos. El que salga 120 unos se nos manifiesta como algo aleatorio pero puede ser la manifestación del  $1/6$  esperado. La estadística nos ayuda a tomar una decisión sobre si es o no el  $1/6$  esperado, y por lo tanto, nos ayuda a decidir sobre si está o no está bien hecho el dado.

Si antes de realizar un estudio, por ejemplo, número de varones y hembras en la familia humana, se afirma que la variable que mide este carácter tiene una variabilidad que se ajusta a una distribución  $Bin(P=0.5)$ , se está aplicando un modelo probabilístico que explica muchos aspectos de este carácter antes de realizarse el estudio, lo que permite prever multitud de resultados y permite enfocar el estudio de una manera correcta.

## **Variables.-**

Los datos en Estadística son observaciones o medidas tomadas sobre la más pequeña unidad muestral, es decir, sobre *unidades experimentales* que no tienen que ser, necesariamente, individuos biológicos.

La característica medida en las unidades experimentales se denomina *carácter* o *variable* y ésta determinará la distribución o población.

Se denomina variable aleatoria porque no se puede predecir el valor que tomará antes de realizar el experimento o realizar la observación, por lo que va implícito el concepto de incertidumbre. El grado de incertidumbre de que una variable aleatoria tome un determinado valor es viene determinado probabilísticamente. Y la probabilidad puede medirse como el número de casos favorables dividido por el número de casos posibles. Por ejemplo, considérese el experimento anterior del dado perfectamente construido, la variable es el número que saldrá cuando se arroje el dado. Esta variable puede tomar cualquier valor entero entre uno y seis, pero es evidente que jamás se sabrá con absoluta certeza cuál de los seis números saldrá cada vez que se tire el dado; se sabe que tiene que ser uno de los seis y se sabe sus probabilidades, es decir, se sabe su distribución, pero no se sabe cuál de ellos.

El conocimiento de la probabilidad de un suceso no elimina la incertidumbre que se pueda tener respecto a su ocurrencia, sino que nos informa sobre su mayor o menor posibilidad de ocurrencia.

## **Clases de variables por su escala de medida.-**

Los tipos de medidas más frecuentemente usadas son: *Nominal* o *de clase*,

*Ordinal*, de *Intervalo* y de *Proporción*. Lo que caracteriza los diferentes tipos de medidas es las relaciones que pueden establecerse entre sus observaciones y, como consecuencia, las operaciones aritmético/lógicas que admiten. El tipo de medida es importante puesto que de él dependerá el espacio muestral, la función de distribución y por lo tanto el método estadístico que se pueda utilizar.

Una medida es Nominal o de clase cuando se le da un nombre al valor de la variable. Por tanto, la única información que provee una observación es a cuál clase pertenece de entre un cierto número de clases o categorías. La única relación que puede establecerse entre observaciones medidas de esta manera es la de igualdad y desigualdad. Dos observaciones serán iguales si pertenecen a la misma clase (tienen el mismo nombre). Se podrá realizar estadística con este tipo de variable porque se tienen las frecuencias de cada clase. Ejemplo de este tipo de medida es la variable *sexo* tiene dos medidas *machos* y *hembra*; podríamos haberlas denominado 1 y 2, lo cual no es sino otra manera de nominarlas, sin ningún valor como números.

Se tiene una medida Ordinal, cuando los valores de este tipo de variable pueden ordenarse de mayor a menor, y en consecuencia no sólo admiten la relación de igualdad. Con las observaciones ordinales se puede calcular frecuencias como con las anteriores, pero además se puede calcular la mediana de una muestra puesto que las observaciones son ordenables. Se dice, entonces, que las variables ordinales son más fuertes (tienen más información) que las nominales o de clase. Un ejemplo de este tipo de variable es el de la clasificación de los sementales por su estampa morfológica o la clasificación de los deportistas en medalla de oro, plata, bronce, etc. Otros ejemplos pueden ser; las clases sociales, que se pueden ordenar en *alta*, *media* y *baja*; los accidentados o enfermos pueden clasificarse en *leve*, *menos grave*, *grave* y *muy grave*.

Las observaciones de variables de Intervalo no sólo pueden ordenarse, sino que además puede definirse una unidad de distancia (si bien arbitraria) entre ellas. Puesto que una unidad de distancia requiere de un cero, es indispensable que exista el cero en la escala, pero tanto el cero como la unidad de distancia son arbitrarios y, en particular, el cero no corresponde a una característica física de las unidades de medida. Un ejemplo clásico de este tipo de variable es la medida de la temperatura donde tanto el cero (cero grados no indica ausencia de calor o de frío) como la unidad de distancia (un grado) son arbitrarios, tanto que pueden cambiar de una convención a otra (*Celsius* o *Fahrenheit*) pero la diferencia con la ordinal es que en las variables de intervalos la unidad de medida es común y constante, asignando un número real a todos los pares de medidas en un conjunto ordenado, mientras que en las variables ordinales las distancias entre variables consecutivas no tienen sentido. En este tipo de variables la proporción de las diferencias entre los valores de una escala es igual a la proporción entre las diferencias correspondientes de la otra; por ejemplo, si estamos en la escala de *Celsius*, la proporción de las diferencias entre 30°C y 10°C, y 10°C y 0°C es  $(30-10)/(10-0)=2$ , y los valores equivalentes en la escala *Fahrenheit* da la proporción  $(86-50)/(50-32)=2$ , mientras que  $30/0 \neq 86/32$  pues tanto el cero como la unidad de medida usada son arbitrarios. Puesto que los requisitos indispensables para efectuar sumas y productos son que existan un cero y una unidad de distancia, con este tipo de variables pueden calcularse la media y la varianza además de los estadísticos de las variables anteriores. Los datos de este tipo de variables tienen, por tanto, más información que los de la variable ordinal, y en ciertos casos se pueden realizar con

ellos las pruebas estadísticas propias de las variables de proporción.

Las variables de Proporción son las variables continuas para las que se usa el modelo normal. En estas variables las observaciones pueden ordenarse y existen un cero y una unidad de distancia que son inherentes al sistema, es decir, que no son arbitrarios. En estos tipos de variables, la proporción de un punto a otro cualquiera es independiente de la unidad de medida; por ejemplo, si pesamos dos objetos en libras y en gramos, la proporción de peso de estos dos objetos es la misma, independientemente de las unidades empleadas, así como el cero que mide falta de peso tanto en libras como en gramos. Los datos de este tipo de variable son los que contienen más información porque son verdaderos números con un verdadero cero, solamente la unidad de medida es arbitraria, por lo que es con este tipo de variables con las que son apropiados los métodos estadísticos más potentes.

Los datos de un tipo de medida siempre pueden transformarse a otra más débil teniendo en cuenta que se está perdiendo información. Y lo mismo se puede decir de las pruebas estadísticas, una prueba estadística de un tipo de variable se puede aplicar a otro tipo de variable con mayor información, pero no se puede aplicar a otra variable de información inferior; es decir, una prueba estadística propia de una variable ordinal se puede aplicar a una variable de clase o de proporción pero no a una variable nominal. Aunque no son recomendables estos cambios por pérdida de precisión.

### **Variable cuantitativa / variable cualitativa.-**

Las variables nominales son *cualitativas*, esto es, distintos niveles o clases difieren en cualidad no en cantidad. Las variables de intervalo y de proporción son *cuantitativas*, esto es, distintos niveles o valores son diferentes en la característica de interés.

Las variables ordinales son de difícil clasificación en lo que concierne a lo cualitativo o cuantitativo, y de hecho hay veces que se utilizan como nominales y, por lo tanto, como cualitativas, y otras veces que se utilizan como cuantitativas; por ejemplo, si los pacientes de cierta enfermedad se ordenan por su gravedad en *leve*, *menos grave* y *grave* no cabe duda de que esta clasificación es una medida de una característica inherentemente cuantitativa.

### **Variable discreta / variable continua.-**

Las variables se pueden clasificar como *continuas* o *discretas* de acuerdo con el número de diferentes valores que pueden tomar. La escala discreta tiene sólo un limitado número de valores, mientras que la escala continua tiene, conceptualmente, un infinito número de valores. En la práctica, muchas variables continuas son medidas en una escala discreta debido a la limitación de los instrumentos de medida.

Una variable es medida en una escala discreta si tiene un número limitado de valores en su campo de variación. Por ejemplo, el número de huevos puestos por una gallina a la semana es una variable medida en escala discreta; simplemente, puede no poner ningún huevo, lo que da el valor de cero, o poner 1, 2 o 7 huevos, lo que dará los

valores de 1, 2 o 7, respectivamente; está claro que una gallina no puede poner 3.5 huevos en una semana.

Una variable es medida en escala continua cuando tiene, conceptualmente, un número ilimitado de valores. Por ejemplo, si se mide la temperatura corporal en humanos se puede obtener uno de los infinitos valores que existe dentro del campo de variación de la temperatura corporal, aunque de hecho son unos pocos los valores que realmente se van a medir, pero no porque no existan, sino porque no tenemos la capacidad de medirlos o para la finalidad del estudio no es necesario más precisión.

Las variables nominales y ordinales se miden, normalmente, en la escala discreta, mientras que las variables de intervalos y de proporciones se miden en escala continua pero se pueden transformar en escala discreta agrupando los individuos en un pequeño número de categorías; por ejemplo, se tiene altura de personas en *cm* y las agrupamos en tres clases: *altas, medianas y pequeñas*.

### **Datos de respuestas categóricas.-**

Una variable categórica es la que representa un número o frecuencia de unidades que caen dentro de una categoría o clase de entre un conjunto de clases. Por ejemplo, la ideología política puede medirse como de *izquierda, centro y derecha*; el estatus de los fumadores puede medirse como *nunca fuma, fuma esporádicamente y fuma asiduamente*; etc.

Estos tipos de datos son muy corrientes en Ciencias Sociales y Ciencias Biomédicas. También se presentan con cierta frecuencia en Etología, Ecología, Sanidad Pública, Ciencias de la Educación, Marketing. Y tiene campos de aplicación muy interesantes en los Controles de Calidad Industrial en los que hay que evaluar características a menudo muy subjetivas como puede ser la característica al tacto de cierto producto, o el buen sabor de un cierto alimento, etc.

### **Variable de Respuesta y variable Explicativa.-**

Muchos análisis estadísticos distinguen entre *Variable Respuesta* (o variable dependiente) y *Variable Explicativa* o (variable independiente). Por ejemplo, los modelos de regresión describen como la distribución de una respuesta (variable) continua (como puede ser tiempo de supervivencia a un trasplante de corazón) cambia de acuerdo a los valores de la variable explicativa, como puede ser edad del paciente o niveles de colesterol del paciente.

### **Exactitud y precisión de los datos.-**

*Exactitud*, es la cercanía de una medida a su verdadero valor. *Precisión*, es la cercanía entre medidas repetidas del mismo individuo.

La precisión es más importante que la exactitud. Por ejemplo, es más importante una balanza de precisión desequilibrada que una vieja y oxidada balanza de

mercado equilibrada, porque la primera medirá siempre la misma diferencia de peso entre dos unidades experimentales de peso próximo, aunque los pesos medidos no sean exactos; mientras que la segunda dará diferentes diferencias de peso entre dichas unidades experimentales, pudiendo dar como más pesado una unidad y en otra pesada dar como más pesado la otra unidad, aunque su peso sea más próximo al exacto que el de la balanza de precisión desequilibrada.

Los valores exactos son usualmente, aunque no necesariamente, de variables discretas y de variables cualitativas (que son derivadas de variables discretas como son las frecuencias). Por lo tanto, el término de *error*, cuando se este tratando estos tipos de variables, no se deberá tanto al error de medida como al error de muestreo y al error de experimentación.

La mayor parte de las variables continuas, sin embargo, son aproximadas. El valor exacto de una única medida es desconocido y probablemente incognoscible, pero el último dígito de la medida realizada debe denotar precisión, es decir, debe marcar con precisión los límites reales sobre la escala de medidas entre los que se está seguro que se encuentra el valor exacto.

### **Ejemplo.-**

Si se afirma que una observación de cierta variable continua es 193, por su carácter de continuo, no se está seguro de esa medida pero si se está seguro que su valor es superior a 192.5 y es inferior a 193.5, si no fuera así, no se podría dar dicho valor a la variable. Si se está usando un instrumento de medida más preciso y se afirma que mide 192.8, no se esta seguro del último 8 pero si se esta seguro de que esa unidad mide más de 192.75 y menos de 192.85, etc.

Si la medida exacta fuera 192.78 unidades, es obvio que la exactitud en estas medidas va en aumento (a la vez que aumenta la precisión) puesto que los límites reales de la medida son mucho más amplios en la primera que en la segunda y más amplios en la segunda que en la tercera.

Para saber el número de cifras significativas que se debe tomar para que los errores de medida cometidos sean no significativos, se ordena la muestra por orden de magnitud, desde la medida más pequeña a la más grande, y el número de unidades que separe a ambas tiene estar comprendido entre 30 y 300.

### **Ejemplo.-**

Medidas en *mm*, la más pequeña 5 *mm* y la mayor 9 *mm*, hay 4 pasos o unidades (9-5), por lo que se tendría que utilizar un instrumento que mida con una cifra significativa más de manera que la más pequeña fuera, por ejemplo, 5.1 y la mayor 9.5, con lo que se tendría ahora 44 pasos (95-51), que es un valor adecuado. La razón para esto es que si se comete un error de una unidad (en el último dígito), en el primer caso se esta cometiendo un error del 25% (1/4), mientras que un error en el último dígito en el segundo caso se esta cometiendo un error inferior al 2.5% (0.1/44).



El último dígito de un número ha de ser significativo, es decir, debe denotar una distancia o separación para la verdadera medida desde la mitad de un paso de unidad anterior hasta la mitad de un paso de unidad posterior (límites reales de la variable). Esto también afecta al cero a la derecha de la coma decimal

$$1.80m \Rightarrow 1.795m - 1.805m$$

## POBLACIÓN, MUESTRA, PARÁMETRO Y ESTADÍSTICO.

### Población.-

*Población o Universo* es el conjunto o grupo de elementos total que está bajo investigación atendiendo a que poseen una o unas características comunes. Lo que caracteriza la población no es el conjunto de individuos sino la variable, es decir, el conjunto de números o medidas que valoran o miden la característica que se están estudiando.

El concepto de población, por tanto, es relativo, depende del objeto de estudio.

### Muestra.-

*Muestra* es una parte de la población que se extrae de ésta para estudiarla y utilizar esta información para hacer inferencias acerca de la población que no puede ser estudiada directamente por ser *infinita* o por ser excesivamente caro o por requerir un tiempo excesivo.

Una muestra tiene que ser representativa de la población si se quiere que las inferencias sean válidas. Para que una muestra sea representativa tiene que cumplir el principio de *aleatoriedad*, es decir, que todos los elementos de la población tienen la misma probabilidad de entrar en la muestra.

### Parámetro de una población.-

Se considera como parámetro de una población cualquier característica que ayude a un mejor conocimiento de dicha población. Pueden ser parámetros, el valor máximo y mínimo que puedan tomar la variable bajo estudio (amplitud, rango o campo de variación), el número de individuos que pueden tomar un cierto valor (probabilidades de las clases), el valor medio que toma la variable (media), el valor central del rango (mediana), el valor más abundante (moda), el grado de homogeneidad o heterogeneidad de la población (varianza, desviación típica, coeficiente de variación), el valor que delimita el 5% de la población, etc. Normalmente, los parámetros suelen ser funciones más o menos simples de los valores que forman la

población. En otras palabras, los parámetros son números que describen un aspecto importante de la distribución o población.

### **Estadísticos muestrales y estimas.-**

De la misma manera que se han definido los parámetros se pueden definir ahora los estadísticos aunque con la diferencia de que mientras en la población los parámetros son números, en la muestra los estadísticos son variables aleatorias ya que al ser función de las observaciones y ser éstas aleatorias, también lo serán las funciones o estadísticos.

Por lo tanto, se denominará como *estadístico* a una medida deducida de una muestra que define una característica de dicha muestra. Tal estadístico puede constituir un fin por sí mismo, como una descripción cuantitativa de un atributo de la muestra, o puede proporcionar (misión de la *inferencia*) una *estima* del parámetro correspondiente de la población como consecuencia de que los parámetros de la población no sólo son desconocidos sino que probablemente sean incognoscibles (poblaciones numéricamente infinitas), por lo que se necesitan los estadísticos muestrales como indicadores o estimas de los parámetros poblacionales.

### **Parámetros de posición o de localización y de dispersión.-**

Se estudiarán, ahora, cuáles son los parámetros que mejor describen la población, es decir, la distribución teórica.

El parámetro que mejor permite conocer una población es, desde luego, la función de densidad, pero tiene el inconveniente de que no es manejable. Por lo tanto se estudiarán los parámetros más útiles.

Los parámetros de *posición* o *localización* son números que nos indican en que posición del eje de abscisas se encuentra localizada la distribución. Los más utilizados son la *media*, *moda* y *mediana*.

Los parámetros de *dispersión* son números que nos indican el grado de heterogeneidad de la distribución, es decir, indica si la distribución es ancha o estrecha. Los más utilizados son la *varianza*, la *desviación típica* y el *coeficiente de variación*.

### **Media.-**

La *media aritmética* o simplemente *media* o *promedio* es la suma de todos los valores dividido por el número de valores. Si las medidas se han agrupado en clases (distribución de frecuencias) sería cada medida o marca de clase por el número de individuos que tienen dicha medida o que contiene dicha clase (*factor de ponderación*) dividido por el número total de individuos que componen la población (suma de las frecuencias de todas las clases)

$$\mu = \frac{\sum X_i}{n} = \sum F_i f_i = \sum_i X_i P_i = \sum_i X_i f(x_i) = E(X)$$

La última expresión es lo que se llama en estadística *esperanza matemática* de  $X$  (esperanza matemática de una distribución teórica) que se puede definir, intuitivamente, como el valor esperado antes de medir una variable, y se asocia, por lo tanto, a la media de la función de densidad.

Sólo existe esperanza matemática de variables aleatorias, porque si no es aleatoria no hay probabilidad, por tanto no hay función de densidad y no se podría definir estadísticamente la media pues al no existir incertidumbre no se espera un valor sino que se está seguros del valor.

Las propiedades de la esperanza matemática son, siendo  $c_1$  y  $c_2$  dos constantes, y siendo  $X$  una variable

$$\begin{aligned} E(c) &= c \\ E(cX) &= cE(X) \\ E(c + X) &= c + E(X) \\ E(c_1 + c_2 X) &= c_1 + c_2 E(X) \end{aligned}$$

### **Moda.-**

Es el valor de mayor frecuencia o el valor que se repite más veces. La moda no tiene porque ser única, puede haber distribuciones unimodales, bimodales, etc.

### **Mediana.-**

Es el valor de la variable que divide a la población en dos partes iguales, es decir, el que deja la mitad de los individuos a un lado de la función de densidad y la otra mitad al otro lado de dicha función. Si esta función de densidad fuera simétrica la mediana, moda y media aritmética coinciden.

### **Media geométrica.-**

La media geométrica de  $n$  valores es la raíz  $n$ -ésima del producto de dichos valores

$$G = \sqrt[n]{\prod_{i=1}^{i=n} X_i}$$

### **Media armónica.-**

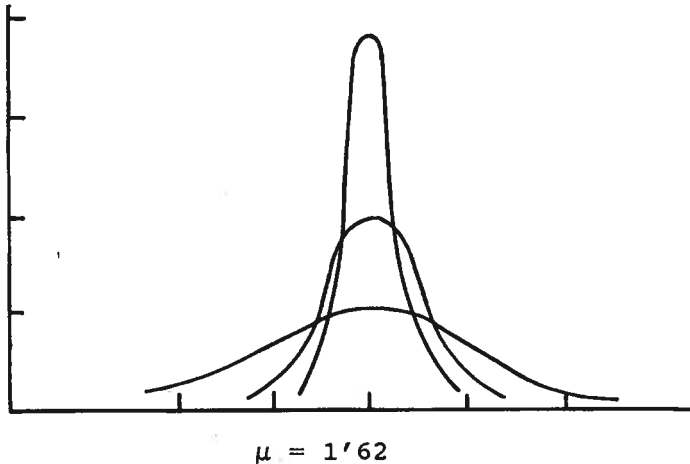
La media armónica de  $n$  valores es la inversa de la media aritmética de la inversa de los valores

$$H = \frac{1}{\frac{\sum_{i=1}^n \frac{1}{X_i}}{n}} = \frac{n}{\sum X_i}$$

### Varianza.-

Supóngase que se tiene la media poblacional ( $\mu=1.62$ ) de la estatura de las españolas, por ejemplo. Ahora interesa saber cómo están agrupadas las españolas por estaturas, es decir, si todas miden 1.62, o bien si hay individuos que miden 1.60 y cuantos son, etc. En definitiva, se quiere saber que forma tiene la función de densidad de esta población que tiene de media 1.62 y que, como se constata, la media, si bien da una información útil, porque localiza la distribución, ésta información es muy limitada.

En la siguiente figura se muestran algunas de las infinitas formas que puede tomar una función de densidad (una población) que tiene 1.62 de media



Uno de los parámetros que podría interesar sería la medida de la variabilidad de esta función de densidad. Para calcularla se podría hallar la desviación de cada medida respecto a la media de la distribución y después sumar todas estas desviaciones; pero esta operación va a dar como resultado el valor de cero, por la misma definición de media, ya que entre estas diferencias se van a compensar las de valor positivo, porque la medida del individuo supera el valor de la media, con las diferencias negativas porque el valor del individuos es inferior al de la media

$$\begin{aligned} \sum_i d_i &= \sum_i (X_i - \mu) = \sum_i X_i - \sum_i \mu = \\ &= \sum_i X_i - N\mu = \sum_i X_i - N\left(\frac{\sum_i X_i}{N}\right) = 0 \end{aligned}$$

Como la suma de estas desviaciones no resuelve el problema, se va a intentar transformarla de manera que mida la variabilidad. Lo primero sería el evitar que las

diferencias se compensen y para ello se pueden elevar al cuadrado para que todas queden positivas

$$\Delta_i = (X_i - \mu)^2$$

Por lo que su suma sería

$$\sum_i \Delta_i = \sum_i (X_i - \mu)^2$$

Pero esto no indica si se refiere a 10, 15, ..., individuos, por lo que habrá que dividir por el número total de individuos y se tendrá la fórmula para el cálculo teórico de la varianza

$$\frac{\sum_i \Delta_i}{N} = \frac{\sum_i (X_i - \mu)^2}{N} = \sigma^2$$

Se denomina *varianza* ( $\sigma^2$ ) en estadística al valor de la variabilidad media de los individuos de la población.

La expresión de la varianza en términos de esperanza matemática es la siguiente

$$\sigma^2 = E[X - E(X)]$$

Las propiedades de la varianza son, siendo  $X$  una variable aleatoria y  $c$  una constante

$$\sigma^2(X) \geq 0$$

$$\sigma^2(c) = 0$$

$$\sigma^2(cX) = c^2 \sigma^2(X)$$

$$\sigma^2(c + X) = \sigma^2(X)$$

$$\sigma^2(X) = E(X^2) - [E(X)]^2$$

### Desviación típica.-

Se representa por  $\sigma$  y es la raíz cuadrada de la varianza. Su utilidad proviene de la necesidad de tener la medida de la variabilidad en las mismas unidades que la variable considerada. Si la unidad de medida de la variable es el metro, la varianza sería en metros cuadrados, por lo que no sería homogéneo con la variable, en cambio la desviación típica sí lo es.

### Coeficiente de variación.-

Para tener una referencia de la importancia de la  $\sigma$  dentro de una población, se tendrá que comparar con algún otro parámetro, pues ella por sí sola no indica si su magnitud es grande o pequeña. Si, por ejemplo, se tiene que  $\sigma = 16.2$  no es lo mismo en una población cuya media sea 162 que en una población cuya media sea 0.162, pues en el primer caso la variabilidad es muy pequeña y en el segundo caso es

enormemente grande.

Para conocer la importancia o magnitud relativa de la  $\sigma$  se referencia con la media y se multiplica por 100 para indicarlo en porcentajes. Este nuevo parámetro se denomina *coeficiente de variación*

$$CV = \left( \frac{\sigma}{\mu} \right) 100$$

Este parámetro permite comparar las variabilidades de dos poblaciones con valores paramétricos diferentes.

### **Consideraciones a tener en cuenta en la elección del análisis apropiado y en la elección de los parámetros apropiados a las variables que se están estudiando.-**

Dos son las causas por las que el decidir estas cuestiones implica cierta dificultad. La primera causa se le puede achacar a los libros de texto que no siempre aclaran suficientemente esta cuestión.

Y la segunda causa es porque los datos que se obtienen en un estudio pueden ser una mezcla de tipos de variables de las vistas anteriormente, o porque son claramente de un tipo pero se han tomado con el propósito claro de realizar un estudio en el que no era preciso la cantidad de información que proveen las variables medidas, o bien porque la escala utilizada invalida la utilización de dicha variable a su nivel de información y hay que utilizarla como una variable de nivel de información inferior.

Otras veces, un mismo conjunto de datos pueden ser analizados de diferentes maneras dependiendo de los supuestos realizados.

Muchas personas, a la hora de realizar un análisis estadístico, pretenden ignorar los supuestos o cuales son los tipos de variables para, dada las facilidades de los paquetes estadísticos, realizar múltiples análisis y elegir el que le provea los resultados más adecuados a las ideas previas. Esta actitud es muy diferente de quien, realizado un análisis teniendo en cuenta los supuestos y los tipos de variables, realiza otro sin tener en cuenta estos supuestos para contrastar los resultados. La primera actitud es muy *peligrosa* pues puede conducir a resultados erróneos si las ideas preconcebidas son erróneas.

## Parámetros adecuados para describir las variables nominales.-

En las variables nominales no tiene sentido el orden de los números utilizados. Por ejemplo, si al sexo masculino se le denomina 1 y al sexo femenino se le denomina 2, esto no implica que el dos sea mayor que el uno, la única relación lógica que se puede hacer es la de igualdad y la de desigualdad, esto quiere decir que todos los individuos denominados como 1 son todos iguales con respecto al sexo y todos ellos son de diferente sexo que los individuos denominados como 2.

La gráfica correcta para la descripción de las variables nominales son las *gráficas de barras* y los *gráficos circulares*. Estos gráficos los realiza, por ejemplo, el SAS con el procedimiento **Chart**. Estos gráficos muestran la proporción de respuesta que tienen los diferentes valores de la variable. Por ejemplo, si se realizan para los sexos se puede visualizar si ambos están en la misma proporción, porque las barras de los dos sexos tengan la misma longitud o los sectores circulares tengan el mismo tamaño, y en caso contrario se ven que no están en la misma proporción.

El parámetro de localización adecuado para estas variable es la *moda* o medida cuyo valor se repite más veces, y el parámetro de dispersión adecuado es la varianza de la distribución a la que se ajuste la variable, esta puede ser la distribución binomial, multinomial o la de *Poisson*. Lógicamente, si esta distribución fuera simétrica, la moda coincidiría con la media.

## Parámetros adecuados para describir las variables ordinales.-

En las variables ordinales si tiene sentido y relevancia el orden de los valores, por lo que son variables con más información que las nominales. Además del gráfico de barras y el gráfico circular, se puede utilizar el *histograma*, aunque el área bajo el histograma no tiene sentido porque los intervalos entre los sucesivos número no son iguales. Por ejemplo, el intervalo entre el primero (medalla de oro) y el segundo (medalla de plata) en una competición atlética no es el mismo que el intervalo entre el segundo y el tercero o entre el cuarto y el quinto; o el intervalo entre el primero y segundo de una oposición no es la misma que entre el segundo y tercero o entre el sexto y el séptimo.

El parámetro de localización adecuado es la *mediana*, este es el valor de la variable que deja a la mitad de los individuos a un lado de la distribución y la otra mitad al otro lado. El parámetro de dispersión puede ser el *rango* o amplitud entre el valor mayor y más pequeño. Otra medida que puede ser informativa en estos tipos de variables es la diferencia entre dos *percentiles*. Por ejemplo, se acostumbra restarle al percentil 95 ( $P_{95}$ ) el percentil 5 ( $P_5$ ). El percentil cinco es el valor de la variable que divide el total de la distribución de manera que el 5% de la población tiene un valor inferior y el 95% restante tiene un valor superior. Hay autores que prefieren el *rango intercuartiles*,  $P_{75}-P_{25}$ , o la *desviación intercuartiles*,  $(P_{75}-P_{25})/2$ . El SAS en el procedimiento **Univariate** provee la mediana, el rango, los cuartiles  $P_{25}$  y  $P_{75}$ , y el rango  $P_{75}-P_{25}$ .

## Parámetros adecuados para describir las variables de intervalo.-

De las variables de intervalo, generalmente se conoce su rango (por ejemplo, la temperatura corporal en °C). La representación adecuada es la de los histogramas, y en este caso si tiene sentido el área delimitada por el histograma, pues el intervalo entre dos valores consecutivos cualesquiera se mantiene constante en todo el campo de variación. Los parámetros de utilidad en estas variables es la media y la varianza.

## Parámetros adecuados para describir las variables de proporción.-

Estas variables son las de mayor información y como son variable que se ajustan a una normal, el parámetro de localización adecuado es la media que coincide con la mediana y la moda, y el parámetro de dispersión adecuado es la varianza. Pero puede haber otros parámetros de utilidad, por ejemplo, si a los datos se le ha realizado la transformación logarítmica es de utilidad la *media geométrica (G)*, puesto que

$$\log G = \frac{\sum \log X}{N}$$

Se usa cuando se quiere saber la media de un proceso en el que hay una proporción constante de cambio; por ejemplo, supóngase un crecimiento rápido de una ciudad que la hecho pasar de 2500 habitantes en 1970 a 5000 habitantes en 1980, para estimar la población de 1975 se puede estimar como

$$G = \sqrt{\left( \prod X_i \right)^{\frac{2}{n}}} = \sqrt{2500 \times 5000} = 3525$$

La *media armónica (H)* se ha definido anteriormente como la inversa de la media aritmética de las inversas de los números. Se usa para obtener una media de proporciones cuando la cantidad en el numerador es fija. Por ejemplo, si un investigador quiere saber la distancia por unidad de tiempo que recorren  $n$  coches en una distancia fija, entonces debe usarse la media armónica.

El coeficiente de variación puede usarse para comparar la variabilidad de distribuciones que tienen diferentes medias (están situadas en coordenadas diferentes).

## Análisis apropiado para los diferentes tipos de variables.-

Para decidir cual es el análisis apropiado es importante tener en cuenta las clasificaciones que se han hecho de las variables tanto por su escala de medida como por su dependencia/independencia.



La clasificación de dependiente o independiente puede ser diferente de un análisis a otro, mientras que la clasificación por su escala de medida permanece constante.

Los análisis apropiados para los diferentes tipos de variables se pueden esquematizar en la tabla de la página siguiente.

En esta tabla se han combinado las variables nominales y ordinales porque no vamos a ver ningún análisis que sea exclusivamente apropiado solo para las variables nominales o solo para las variables ordinales.

También se ha combinado las variables de intervalo y de proporción porque se van a utilizar los mismo análisis para ambos tipos

**Variabes Independientes**

		<b>Nominales y Ordinales</b>		<b>Intervalo y Proporción</b>		
		1 variable	>1 variable	1 variable	>1 variable	
<b>sin variable dependiente</b>		Bondad de ajuste ( $\chi^2$ ) Correspondencias Modelos categóricos	Log-lineal $\chi^2$ independencia Correspondencias Modelos categóricos	Normalidad $t$ una muestra transformaciones	Componentes Principales Análisis Factorial Dendogramas (Cluster)	
	<b>Nominal Y Ordinal</b>	1 variable  >1 variable	$\chi^2$ independencia Prueba exacta de Fisher Modelo log-lineal	Modelo Log-lineal regresión logística	$t$ dos muestras Función discriminante regresión logística	Función discriminante Regresión logística
<b>Variabes Dependientes</b>	<b>Intervalo Y Proporción</b>	1 variable  >1 variable	Modelo log-lineal  $t$ datos emparejados Análisis de Varianza $T^2$ de Hotelling A. Multivariante de la Varianza ANOVA de los C. Principales	Modelo log-lineal  A. Multivariante de la Varianza	Función discriminante Regresión lineal regresión no lineal Correlación	Función Discriminante Regresión Múltiple Correlación Parcial Correlación Múltiple
						Correlación Canónica

## Análisis Univariantes/Multivariante.-

La investigación aplicada, tanto en Ciencias de la Naturaleza como en otras ramas de la Ciencia, frecuentemente se encuentra con fenómenos complejos que requieren para su análisis de una considerable cantidad de variables medidas sobre el mismo individuo estadístico o unidad experimental. Se entiende por **análisis multivariante** la rama de la estadística y del análisis de datos que estudia, interpreta y elabora el material estadístico sobre la base de un conjunto de  $p > 1$  variables dependientes o interdependientes, que pueden ser de tipo cuantitativo, cualitativo o una mezcla de ambas. La información en el análisis multivariante es, por lo tanto, de carácter multidimensional. En general, las variables se asimilan a variables aleatorias con una cierta distribución multivariante. Quizás la propiedad más determinante del análisis multivariante es que las  $p$  variables son dependientes (son función de), de naturaleza similar y que ninguna de ellas tiene una importancia superior a las demás.

En cierto sentido, se puede decir que no existe una frontera clara entre los métodos multivariantes y las demás técnicas estadísticas. La estadística univariante consiste, esencialmente, en describir una sola variable mediante características muestrales, o realizar inferencias en base a la información de una muestra obtenida bajo ciertas condiciones experimentales. El análisis multivariante cumple también con estos dos objetivos, *descriptivo* e *inductivo*, pero trabajando simultáneamente con varias variables en lugar de una sola. Es una metodología estadística **operativamente más complicada**, pero también más potente, que utiliza extensamente los métodos del álgebra lineal, cálculo numérico, geometría lineal, etc., aunque **conceptualmente sea lo mismo** que los métodos univariantes.

Los métodos del análisis multivariante permanecieron en el campo teórico más tiempo que los métodos estadísticos univariantes, como consecuencia de lo engorroso y complicado de su uso práctico. A partir de la década de los setenta, con la generalización del uso de los ordenadores, muchas investigaciones aplicadas utilizan métodos multivariantes que vienen justificados por la multicausalidad de los fenómenos naturales, lo que hace que estos métodos puedan jugar un papel importante en el análisis de datos en las investigaciones de carácter empírico-experimental.

## Tipos de análisis.-

Como se ha visto en la tabla anterior, la cantidad y variedad de los métodos uni y multivariantes pueden producir cierta confusión. Desde la perspectiva de la investigación aplicada, interesa fundamentalmente conocer a cuales problemas es adecuado aplicar cada una de las técnicas, cómo hacerlo y cómo interpretar los resultados. Siguiendo a *Dagnelie* se hará una introducción previa a los problemas agrupándolos, para facilitar la exposición, en cinco categorías, sin pretender que estas categorías sean absolutas ni excluyentes.

## Pruebas de homogeneidad.-

En estos análisis se tienen varias poblaciones y una sola variable o un solo grupo de variables. Las pruebas de homogeneidad de varias variables pueden ser consideradas como simples generalizaciones de las pruebas de homogeneidad clásicas. Así como la prueba *t* de *Student* permite contrastar la hipótesis de igualdad de medias de una variable medida en dos poblaciones, la prueba *T* de *Hotelling* puede servir para verificar la hipótesis de igualdad de medias de dos poblaciones para varias variables, simultáneamente. Si existen más de dos poblaciones (también para dos poblaciones) se tiene el **análisis de la varianza**, con una sola variable y, con dos o más variables, el **análisis multivariante de la varianza**, que conceptualmente es como el análisis de la varianza.

Lo mismo se puede decir de las pruebas de igualdad de varianzas, de igualdad de coeficientes de correlación, etc.

Estas pruebas se estudian en los Capítulos 4 al 10, y en el Capítulo 17 se estudia el Análisis de Covarianza, que se utiliza para pruebas de homogeneidad pero que hace uso de los métodos del análisis de la varianza y del análisis de regresión.

## Regresión.-

En estos análisis se tendrá una población y dos grupos de variables, cada grupo de naturaleza diferente. El caso más simple de regresión es aquel donde se considera dos variables con el fin de prever o estimar el o los valores de una variable, denominada *dependiente* o de *respuesta*, correspondiente a uno o varios valores conocidos de otra variable, denominada *independiente* o *explicativa*. En este caso se utiliza una ecuación de regresión simple, corrientemente será una ecuación lineal y por tanto representativa de una línea recta, determinada a partir de los valores observados para las dos variables en un cierto número de individuos. Tal es el caso, por ejemplo, si se observa durante *n* campañas cierto carácter productivo en una comarca dada y el valor de una cierta característica climática (temperatura máxima, por ejemplo) en el periodo de campaña, y si se desea prever el rendimiento de este carácter productivo en el curso de la (*n*+1)-ésima campaña, en función del valor de la característica climática observada en la (*n*+1)-ésima estación en la que se desarrolla la campaña.

Este tipo de problemas no podemos clasificarlo como de análisis multivariante, pero dos de sus generalizaciones pueden ser consideradas como análisis multivariante. La primera generalización es la de la regresión múltiple. La regresión de *Y* en una sola variable independiente es muchas veces inadecuada. A menudo, puede haber disponibles dos o más *X* para proporcionar más información acerca de *Y*, por medio de la regresión múltiple en las *X*. La finalidad de la regresión múltiple, como ya se ha apuntado, es la de estimar o prever el o los valores de una variable dependiente en función no de una variable sino de varias variables independientes, lo que permite una mejor estima/predicción de los valores de *Y*. Se utiliza una ecuación de regresión múltiple, en general lineal, y por lo tanto representativa de un *plano de regresión* si tenemos dos variables independientes, un *espacio de regresión* si tenemos tres variables y un *hiperespacio de regresión* para más de tres variables independientes.

Una segunda generalización de la regresión simple tiene como objetivo el estudio de las relaciones existentes entre dos grupos de variables con el fin de prever los valores de las variables de un grupo en función de los valores de las variables del otro grupo. Se utiliza con este fin *sistemas de ecuaciones de regresión simultáneas* a menudo llamadas *modelos econométricos*. En este caso, a las variables dependientes se les denomina variables *endógenas* y a las variables independientes, variables *exógenas*.

Si designamos como  $p$  el número de las variables medidas y como  $n$  el número de individuos, tendremos  $np$  valores que en el caso de la regresión múltiple se trata de  $p$  variables independientes, de una variable dependiente y de  $n+1$  individuos ( $n$  individuos completamente conocidos y un individuo en el que se debe estimar el valor de la variable dependiente). Y en el caso de las ecuaciones simultáneas serán  $p$  variables independientes,  $q$  variables dependientes y  $n+1$  individuos ( $n$  individuos completamente conocidos y un individuo en los que se debe estimar el valor de las  $q$  variables dependientes).

Otro uso de la regresión múltiple es la de, si se tienen muchas  $X$  se puede buscar el subconjunto de  $X$  que da la mejor ecuación lineal. Por ejemplo, en la predicción del tiempo se puede usar hasta 50 variables  $X$ ; una ecuación con 50 variables  $X$  puede resultar difícil de manejar, además de no ser preciso el utilizarlas todas si muchas de estas variables no contribuyen significativamente a un mejor pronóstico, por lo que se puede buscar una ecuación con solo las tres o cuatro variables más predictivas.

En otros estudios semejantes al anterior, el objetivo no es el pronóstico, sino que se tiene un número de variables  $X$  y se quiere saber cuáles están relacionadas con la  $Y$ , y, de ser posible, ordenar las variables en categorías según su importancia en influenciar a  $Y$ .

Estas pruebas se estudian en los Capítulos 11 al 14.

## **Correlación.-**

En este caso también se tendrá una población y dos grupos de variables, cada grupo de naturaleza diferente. El *coeficiente de correlación simple*, mide la intensidad de relación lineal existente entre dos variables.

El coeficiente de *correlación múltiple*, tiene un uso semejante al coeficiente de regresión múltiple. Puede usarse como medida de la calidad de la previsión obtenida para una variable en función de dos o más variables.

El coeficiente de *correlación parcial* permite la medida de la intensidad de variación conjunta de dos variables haciendo abstracción de la posible influencia de otra u otras variables. El coeficiente de correlación parcial puede ser utilizado en el curso de un estudio de regresión múltiple para la elección correcta de las variables independientes.

El análisis de *correlación canónica* mide la intensidad de variación conjunta existente entre dos grupos de variables. Es una generalización del coeficiente de correlación múltiple.

En definitiva, los problemas de correlación y regresión están estrechamente ligados y a menudo es difícil de disociar.

Estas pruebas se estudian en los capítulos 15 y 16.

### **Análisis de Componentes Principales y Análisis Factorial.-**

En estos análisis hay una población y un solo grupo de variables, es decir, las observaciones proceden de una misma población y las variables son de naturaleza homogénea, esto es, no hay variables dependientes y variables independientes como ocurre con la regresión.

El análisis de componentes principales se usa para la interpretación de las relaciones existentes entre una serie de variables interdependientes. Su objetivo principal es el de condensar la información que nos provee estas variables interdependientes, observables directamente, en un número más restringido de nuevas variables no correlacionadas, por lo tanto, independientes, que no se pueden observar directamente, denominadas *componentes principales*. Cada componente principal es una combinación lineal de las variables originales.

Si se realiza una regresión múltiple utilizando como variables independientes las componentes principales de las variables independientes originales, se tiene el análisis conocido como *regresión ortogonal*. Este análisis es muy útil sobre todo cuando las variables independientes están fuertemente correlacionadas entre ellas.

El análisis factorial persigue un objetivo análogo al del análisis de componentes, pero siguiendo un modelo matemático algo diferente. El análisis factorial intenta explicar un grupo extenso de variables directamente observables mediante un número reducido de variables hipotéticas llamadas *factores* que no son directamente observables, pues son variables de naturaleza más abstracta que las variables originales. Estos factores y su número son conocidos previamente al análisis.

Estas pruebas se estudian en el Capítulo 18.

### **Problemas de Clasificación.-**

De otra parte, cuando se consideran dos o más poblaciones, entre las cuales existen ciertas diferencias de medias, el problema puede ser el de situar en una u otra población a los individuos nuevos que no se conoce a priori su pertenencia. Este problema de clasificación puede ser resuelto en ciertas condiciones por medio de una función lineal de las características observadas, llamada *función discriminante*. Si de lo que se trata es de analizar un conjunto de observaciones con objeto de verificar si pertenecen a grupos diferentes, se tiene el *análisis factorial discriminante*. También se utiliza el análisis discriminante, como ocurría con la regresión, para encontrar la

variable o variables que mejor discriminan grupos preestablecidos. Si lo que se pretende es buscar grupos similares de individuos o variables se tiene el *análisis o algoritmos Cluster*.

Del problema de clasificación surge todo un completo conjunto de análisis multivariantes denominados genéricamente *Taxonomía numérica*.

Estas pruebas se estudian en el Capítulo 19.

### **Datos categóricos.-**

Con los datos de respuesta categórica son los que representan un número o frecuencia de unidades que caen dentro de una categoría o clase de entre un conjunto de clases. Con estos datos se podrán hacer pruebas de homogeneidad pruebas semejantes a la regresión por medio de los modelos lineales categóricos y pruebas semejantes al análisis de Componentes Principales por medio del Análisis de Correspondencias.

Estas pruebas se estudian en el Capítulo 20.

### **Bibliografía:**

- Afifi, A.A. y Clark, V.* 1984. COMPUTER-AIDED MULIVARIATE ANALYSIS. Ed: Lifetime Learning Publications. Belmont (USA).
- Cuadras, C.M.* 1981. MÉTODOS DE ANÁLISIS MULTIVARIANTE. Ed:EUNIBAR. Barcelona (España).
- Dagnelie, P.* 1982. ANALYSE STATISTIQUE À PLUSIEURS VARIABLES. Ed: Les Presses Agronomiques De Gembloux. Gembloux (Belgique).
- Dixon, J.D. y Manssey, F.J.* 1974. INTRODUCCIÓN AL ANÁLISIS ESTADÍSTICO. Ed. Del Castillo. Madrid.
- Infante Gil, S. y Zárate De Lara, G.P.* 1984. MÉTODOS ESTADÍSTICOS. Ed. TRILLAS. México.
- Gilbert, N.* 1976. ESTADÍSTICA. Ed. Interamericana. México.
- Lebart, L., Morineau, A. Y Fénelon, J.P.* 1979. TRAITEMENT DES DONNÉES STATISTIQUES. Ed: Dunod. Paris (France).
- Lefebvre, J.* 1980. INRODUCTION AUX ANALYSES STATISTIQUES MULTIDIMENSIONNELLES. Ed: Masson. Paris (France)
- Littel, R.C., Freund, R.J. and Spector, P.C.* 1991. SAS<sup>®</sup> FOR LINEAR MODELS. SAS Institute Inc., Cary, NC, USA.
- Milton, J.S.* 1994. ESTADÍSTICA PARA BIOLOGÍA Y CIENCIAS DE LA SALUD. Ed. Interamericana-McGraw-Hill. México.
- Sokal, R.R. y Rohlf, F.J.* 1994. BIOMETRY. Ed. W.H.FREEMAN. San Francisco.
- Srivastava, M.S. y Carter, E.M.*1983. AN INTRODUCTION TO APPLIED MULTIVARIATE STATISTICS. Ed:Elsevier Science Publishing. New York (USA).
- Steel, R.* 1996. PRINCIPLES AND PROCEDURES OF STATISTICS. Ed. McGRAW-HILL Education. New York .

## **CAPÍTULO 3**

# **Pruebas de hipótesis**





## Pruebas de hipótesis

El objetivo último de la estadística es el de ayudarnos a tomar decisiones sobre la estructura del modelo usando los datos, es decir, inferir sobre la población usando la muestra que se tiene. Y esto se hace mediante una *Prueba de Hipótesis*. Expliquemos qué es esto.

El conocimiento científico consiste, en su primer paso, en la formulación de un *problema* planteado en una población; un problema que se presente una sola (o pocas) vez no es un problema científico, esto es, tiene que haber repetibilidad tanto en el problema como en la solución que se le dé. Por lo que el problema surge como consecuencia de la observación de unos hechos y el objetivo es el entenderlo, interpretarlo y, por tanto, poder incidir sobre él transformando, de esta manera, la realidad.

Si la ciencia en que se presenta el problema está en los primeros estadios de su desarrollo, las soluciones de los problemas suelen ser descriptivas, pero enseguida que se asientan estas primeras descripciones comienzan a surgir problemas que requieren cuantificaciones e inferencias. Pongamos cinco problemas a modo de ejemplos representativos.

- 1) Un problema puede ser, por ejemplo, el planteado con el objetivo de saber si existe presión sanguínea en las venas.
- 2) Otro problema puede ser el planteado a la hora de medir el área de tres superficies de terreno, perfectamente cuadradas, adjuntas de manera que uno de los extremos de sendos lados de las dos superficies más pequeñas se tocan perpendicularmente, mientras que los otros extremos de dichos lados se tocan con sendos extremos de un lado de la superficie mayor.
- 3) Otro problema puede ser el que se plantea cuando se evalúa los conocimientos de la mente humana en una materia. Por ejemplo ¿Conoce un alumno concreto la materia de la asignatura Estadística?

- 4) Otro problema puede ser el de, si en cierta especie de vertebrados, el número de nacimientos de machos es igual al de hembras.
- 5) Otro problema puede ser el de, si en cierta especie de mamífero domestico, el tamaño de camada es 3.
- 6) Otro problema se plantea cuando se observa que ciertas características del hombre o de otras especies, vitalmente desfavorables (por ejemplo, el albinismo), aparece en una generación, desaparece en la siguiente, y puede volver a aparecer en los nietos o en otra generación futura.

Una vez planteado un problema se le busca una solución. La solución puede ser de dos tipos: solución descriptiva o solución por medio de una hipótesis contrastable. Las soluciones descriptivas no hacen sino plantear inmediatamente más problemas que requieren para su solución de hipótesis contrastables.

La solución al primer problema es descriptiva, esto es, la de realizar observaciones y mediciones que informarán y proveerán la solución de si existe presión sanguínea en las venas. Esta primera fase descriptiva es común a todas las ciencias y es la que provee de elementos para la elaboración de hipótesis para resolución de nuevos y más complicados problemas, como puede ser, ¿Porqué cuando se aplica un determinado producto químico aumenta la presión sanguínea?

Los demás problemas no pueden ser resueltos descriptivamente, pues las cuestiones que plantean son *inobservables* (no se puede ver lo que ocurre en la mente humana, no se puede ver todos los pasos que va desde el *ADN* hasta la producción del albinismo, no se puede ver todos los procesos citológico, fisiológicos, embriológicos, *etc.*, que va desde la meiosis hasta el nacimiento del individuo, *etc.*). Sin embargo, si se ven sus manifestaciones, los hechos reales, por lo que se pueden resolver estos problemas con la elaboración de *hipótesis*. Esto es, para resolver un problema de este tipo se emite una idea provisional de cómo se ha de interpretar los hechos que han propiciado el problema o cual es el *factor* o cuales son los factores que han dado lugar al hecho observado. Esta idea provisional, por tanto, es una hipótesis contrastable con los hechos reales, lo que dará lugar, u orientará, a una serie de observaciones o experimentos para la obtención de datos relevantes al problema. Con la hipótesis no solo se podrá explicar unos hechos, sino que se podrán prever otros hechos aun no observados y que puede ponerse de manifiesto experimentalmente.

Pero, aunque las hipótesis permiten poner de manifiesto factores inobservables, al ser los fenómenos naturales multicausales, esta manifestación de los hechos nunca es determinante. Y como nunca se puede estudiar la población, hay que tomar una decisión, con respecto a la falsedad o no de la hipótesis en la población, en base a hechos, observados en una muestra, no determinantes.

Expliquemos lo que quiere decir hechos no determinantes con un ejemplo sencillo. Se puede afirmar que si se lanza un dado, la probabilidad de que salga el uno es  $1/6$ ; la única manera determinante que se tendría de confirmar esta afirmación sería observando la población de resultados del lanzamiento de un dado, esto es, lanzando un dado infinitas veces; como esto no es posible, hay que tomar una muestra para decidir sobre la falsedad o no de la afirmación en la población; por ejemplo, se puede

lanzar el dado 600 veces; si las 600 veces sale el uno (o cualquier otro número) se puede rechazar la hipótesis como falsa; si de las 600 veces que se lanza el dado sale 100 veces el uno, se puede aceptar que, al menos por el momento, no es falsa, pero las manifestaciones naturales nunca son tan claras que sea suficiente el sentido común para tomar una decisión. Probablemente saldrá el uno un número de veces entre el 100 y el 600, por lo que en base a ese resultado hay que decidir sobre la falsedad o no de la hipótesis.

Para el segundo problema se puede elaborar una hipótesis que puede ser: *la suma de los cuadrados de los dos lados menores de un triángulo rectángulo es igual al cuadrado del lado mayor*. Como se ha dicho anteriormente, esta hipótesis se podría confirmar si se miden los lados de los infinitos triángulos rectángulos y se confirma que la suma de los cuadrados de los catetos es igual al cuadrado de la hipotenusa. Como esto no puede ser, se puede tomar una muestra de triángulos rectángulos y ver si se confirma esto. Sin embargo, los matemáticos lo demuestran sin medir ni un solo triángulo rectángulo, es decir, esta hipótesis puede demostrarse analíticamente y aceptar que es cierta para todos (la población) los triángulos rectángulos. Pero las deducciones deductivas solo son posibles en las Matemáticas y ciencias afines.

En las ciencias empíricas o *factuales* hay que proponer hipótesis o teorías concernientes a algún universo o población y la única forma de tener una certeza absoluta de su veracidad o falsedad es la de examinar toda la población y como esto es, no solo difícil, sino imposible, es por lo que la solución es tomar una muestra de la población y usar las observaciones de la muestra para tomar una decisión sobre la falsedad o no de la hipótesis.

Lo expresado hasta aquí concierne a todas las ciencias empíricas y a todas las técnicas. Todos los conocimientos que se tienen de fisiología, bioquímica, patología, etc., están basados en estudios realizados en muestras extraídas de una población, a efectos, infinita, con el fin de confirmar o no una hipótesis planteada para la resolución de un problema.

Por tanto, en todas las ciencias empíricas y en todos los campos técnicos se plantean problemas y se elaboran hipótesis que hay que contrastar. En la contratación de dichas hipótesis es en la que interviene los *Métodos estadísticos* como herramientas que permiten realizar *Pruebas de Hipótesis*, esto es, decidir sobre la falsedad o no de una hipótesis.

Para poder probar las hipótesis hay que realizar varios pasos previos.

- 1) Lo primero es decidir cual es la variable que puede cuantificar el carácter y poner de manifiesto el factor expresado en la hipótesis, que solo se podrá observar a partir de sus manifestaciones.

Por ejemplo, para el primer problema (o derivados de este), hay que decidir como se mide la presión sanguínea, esto es, cuál es el aparato o instrumento de medida que pondrá de manifiesto la presión sanguínea, y cuál es la precisión que se utiliza. Se puede decidir realizar las medidas con un manómetro de mercurio y con una precisión lo suficiente alta como para que haya unas 50 unidades entre el máximo y el mínimo del campo de variación. Para el problema tercero, hay que hacer lo mismo y es

obvio que será una variable y unos aparatos muy diferentes a los del primer problema. se puede decidir si un individuo tiene o no ciertos conocimientos midiendo si acierta más veces, de las que sería de esperar por el azar, experimentos aleatorios independientes en los que hay dos posibles resultados mutuamente excluyentes.

2) Lo segundo es decidir cual es el modelo que explica la variabilidad de la variable elegida, es decir, cual es la población de esta variable.

Por ejemplo, en los derivados del primer problema, dado la variable que se ha decidido que se va a utilizar para medir el carácter presión sanguínea y dado la precisión que se va a utilizar, se acepta que esta variable tiene una distribución normal. La variable que se ha decidido utilizar para poner de manifiesto el posible conocimiento en una materia, del tercer problema, se asume que es una variable binomial.

3) El tercer paso es traducir la hipótesis planteada al lenguaje estadístico, lo que implica la elección del estadístico de prueba (ver más adelante).

Este tercer paso y los siguientes, se explicarán más adelante.

Como se ve, en su contexto más general, una hipótesis estadística es una afirmación sobre un modelo probabilístico (población) que una vez asumido, la única o únicas constantes desconocidas son los parámetros de la distribución (población) correspondiente. Aunque es necesario destacar el hecho de que no todas las hipótesis estadísticas se refieren necesariamente a un parámetro de una distribución concreta, hay hipótesis estadísticas no paramétricas, como se verá en el capítulo 20 de **Datos Categóricos**.

Y una prueba de hipótesis es un método para dictaminar sobre la probabilidad de esa afirmación, usando muestras como instrumento. Ya se ha utilizado el ejemplo de la probabilidad de  $1/6$  para el uno, como hipótesis de lo que puede ocurrir al lanzar un dado, y la prueba puede ser la de lanzar un dado 600 (muestra) veces; si el resultado de dicho muestreo provee de 600 unos es fácil, intuitivamente, desechar la hipótesis de que la probabilidad de obtener un uno es  $1/6$ ; por otra parte, si se obtuvieran 100 unos, se aceptaría la hipótesis. Con resultados como estos, sería suficiente con el sentido común para aceptar o rechazar una hipótesis; pero la realidad cotidiana de los experimentos no conduce nunca a resultados tan explícitos, de aquí la necesidad de realizar pruebas estadísticas.

Una vez realizados los dos pasos anteriores, se tomará la muestra, esto es, se realizará las observaciones o los experimentos que nos proveerán de la manifestación real de lo expresado en la hipótesis. Si las observaciones concuerdan con las predicciones expresadas en la hipótesis, se concluye que la prueba no rechaza, por el momento, la hipótesis. Y se rechazará la hipótesis si las observaciones no concuerdan con la misma.

Para que una hipótesis pueda ser rechazada o no rechazada, esta claro que debe ser coherente con solo algunos de los resultados u observaciones posibles. Esto es, las predicciones derivadas de una hipótesis han de ser suficientemente exactas como para que delimiten la región de observaciones posibles en la que la hipótesis

resulte compatible o no falsa. Y debe estar claro y ser exacto la región de posibles resultados que harán rechazar la hipótesis por falsa.

Por lo que para probar una hipótesis estadística se divide el espacio de valores posibles del parámetro, en el que se está interesado, en dos subconjuntos; uno de ellos es el espacio definido por la *hipótesis nula* que constituirá la *región de aceptación* y el otro espacio es el definido por la *hipótesis alternativa* que constituirá la *región de rechazo* (ver más adelante).

Una vez tomada la muestra se decide que, con ciertas probabilidades de error, el parámetro en cuestión pertenece a uno de esos dos subconjuntos. Pero no existe certeza de que no se cometerá un error.

## Error Tipo I, Error Tipo II.-

Para presentar las ideas centrales de una prueba de hipótesis utilizaremos el ejemplo de un examen en el que se le plantea al alumno una pregunta y se le proponen cinco respuestas de las cuales una es la correcta. Esto responde al problema tercero de los planteados anteriormente.

Supóngase un alumno (*A*) que dice conoce la materia, es decir, conocerá la respuesta de la pregunta un número mayor de veces que el que cabría esperar por puro azar. Está claro que solo hay dos resultados posibles: que el alumno acierte o que no acierte. En el caso de que acierte podría apoyar el supuesto de que el alumno conoce la materia, mientras que si no acierta podría ir en contra de esta aseveración.

Una manera de mejorar la capacidad de decisión es repetir el experimento varias veces, puesto que es claro que si el alumno acierta una proporción alta en un gran número de preguntas, su aseveración estará mejor apoyada que si acierta un sólo intento. Supóngase, pues, que el experimento descrito se repitiera 10 veces, esto es, un examen tipo *test* de 10 preguntas con cinco posibles respuestas cada una. (10 es una muestra muy pequeña, pero se toma el valor de  $n = 10$  para que quepa todo el desarrollo del ejemplo en un espacio razonable de este manual). Se tendrá, por lo tanto, que la variable que se registra es el número de veces que el alumno acierta.

Tal como se ha planteado hasta el momento, el experimento no se presta a conclusiones de tipo estadístico, debido que hasta el momento se está manejando un modelo verbal de la realidad. Por lo que el siguiente paso es traducir esta formulación y generar un modelo probabilístico. Está claro que en este caso se tiene un experimento que se ajusta a una *Binomial* donde, si se llama éxito a un acierto de *A*, el parámetro de la distribución es  $p$ , la probabilidad de un acierto. Por lo que se tiene que decidir sobre la magnitud de  $p$ , para ello se puede adoptar la decisión más simple, como es que *A* no conoce la materia y simplemente trata de adivinar la respuesta por puro azar. Traducido al lenguaje probabilístico, se tiene una función de probabilidad *binomial* con  $p=1/5=0.2$ ; mientras que *A* pretende que  $p$  es mayor de 0.2. Se tiene, entonces, una partición natural del espacio de posibles valores para  $p$ . Los valores de  $p \leq 0.2$  corresponden a lo que se podría llamar nuestra hipótesis, mientras que los valores  $p > 0.2$  corresponden a la hipótesis de *A*. Esta partición del espacio de los valores del parámetro en dos conjuntos mutuamente excluyentes es un aspecto importante de la

prueba de hipótesis, como es que a toda hipótesis le corresponde otra complementaria que es su negación. A la primera hipótesis se le denomina *Hipótesis Nula* ( $H_0$ ) que es aquella que el científico ha decidido contrastar y a su complemento o negación se le denomina *Hipótesis Alternativa* ( $H_1$  o  $H_a$ ). Estas son

$$H_0 : p \leq 0.2$$

$$H_1 : p > 0.2$$

La distribución de probabilidades de  $X$  depende, por tanto, del valor de  $p$  o, dicho de otra forma, del conocimiento que tiene  $A$  de la materia. Y sólo se estaría dispuesto a rechazar la  $H_0$  si  $A$  acierta más veces que las que acertaría cualquier persona diciendo al azar un número del 1 a 5.

Como se ve, ante estas dos hipótesis puede ocurrir cualquiera de las dos, ya que un individuo sin conocimiento puede acertar un número suficiente de preguntas y un individuo con conocimiento puede fallar. Por lo tanto, en el momento de tomar una decisión se puede estar errando de dos maneras: decidiendo que  $A$  conoce la materia cuando no la conoce y decidiendo que  $A$  no conoce la materia cuando si la conoce.

Estadísticamente: se puede rechazar la  $H_0$  cuando es cierta y se puede no rechazar la  $H_0$  cuando es falsa.

		Decisión tomada	
		$H_0$ se acepta	$H_0$ se rechaza
Situación real (desconocida)	$H_0$ cierta	No Error	Error Tipo I $\alpha$
	$H_0$ falsa	Error Tipo II $\beta$	No Error

$$\alpha = P(\text{Error Tipo I}) = P(\text{rechazar la } H_0 \text{ cuando es cierta})$$

$$\beta = P(\text{Error Tipo II}) = P(\text{aceptar la } H_0 \text{ cuando es falsa})$$

En el caso de  $A$ , se estará cometiendo un Error Tipo I si se decide que conoce la materia (está aprobado) cuando no la conoce (debiera estar suspenso) y se cometería un Error Tipo II si se decide que no la conoce (está suspenso) cuando si la conoce (debiera estar aprobado).

### Función de potencia de una hipótesis nula.-

Puesto que lo que se desea es cuantificar la evidencia o incertidumbre para rechazar  $H_0$ , un concepto importante en una prueba de hipótesis es la probabilidad de rechazar una hipótesis dada. Puesto que esta probabilidad es una función del parámetro que se esta probando, se simboliza por  $\delta(p)$

$$\delta(p) = P(\text{rechazar } H_0 \text{ cuando el verdadero valor del parámetro es } p)$$

Por lo dicho anteriormente, se ve que para todos los valores de  $p$  en la  $H_0$

$$\delta_{(p)} = \text{Probabilidad de Error Tipo II}$$

Mientras que para todos los valores de  $p$  en la  $H_1$

$$\delta_{(p)} = 1 - \text{Probabilidad de Error Tipo II}$$

Puesto que  $\delta_{(p)}$  es la probabilidad de rechazar una hipótesis dada, dependiendo del verdadero valor del parámetro, se la llama *función de potencia de la prueba*.

### Estadístico de prueba y regla de decisión.-

El estadístico para una prueba de hipótesis es el de la variable aleatoria cuyo valor, en una muestra, determinará la decisión de rechazar o no la hipótesis nula. De acuerdo con la distribución de esta variable aleatoria se elegirá una regla de decisión, que es simplemente una regla que establece cuales valores del estadístico producirán el rechazo de  $H_0$  y cuales la aceptación.

El conjunto de valores, en el espacio muestral del estadístico, para los cuales se rechazará  $H_0$  es llamado *Región de Rechazo* de la prueba. Por tanto, una regla de decisión genera una partición del espacio muestral del estadístico de prueba en dos conjuntos mutuamente excluyentes. Uno de ellos es la *Región de Rechazo* de  $H_0$  o *valores críticos* y el otro la *Región de no Rechazo* (o aceptación) de  $H_0$ .

En el ejemplo del alumno que pretende tener ciertos conocimientos, puesto que el experimento es muy sencillo, las reglas de decisión posibles son bien fáciles. La variable aleatoria  $X_i$  sólo puede tomar dos valores; se dirá que  $X_i$  toma el valor uno si  $A$  acierta y toma el valor cero si  $A$  no acierta. Se tiene, entonces, que la distribución de probabilidades de la variable aleatoria  $X_i$  depende del valor de  $p$  o, dicho de otra forma, de la capacidad de conocimiento de  $A$ . Se tiene entonces

$$P_{(X=1)} = p$$

$$P_{(X=0)} = 1 - p = q$$

Por lo que solo se estaría dispuesto a rechazar la  $H_0$  si  $A$  acierta más veces que las que acertaría cualquier persona diciendo al azar un número del 1 a 5.

Para analizar el comportamiento de los Errores, de Tipo I y de Tipo II de estas hipótesis se analizaran estas tres posibles reglas de decisión

<i>Prueba</i>	<i>Regla de decisión</i>	<i>Región de Rechazo</i>
1	<i>Rechazar <math>H_0</math> si <math>X \geq 3</math></i>	$R_1 = (X   X=3,4,5,6,7,8,9,10)$
2	<i>Rechazar <math>H_0</math> si <math>X \geq 5</math></i>	$R_2 = (X   X=5,6,7,8,9,10)$
3	<i>Rechazar <math>H_0</math> si <math>X \geq 8</math></i>	$R_3 = (X   X=8,9,10)$



Cada una de las tres reglas se caracterizarán por una región de rechazo diferente. El estadístico para la prueba será el de la variable  $X$  distribuido  $Bin(10, p)$ . Es decir, que  $X$  será el número de veces que  $A$  acierte de un total de 10 intentos.

*Prueba 1) Rechazar  $H_0$  si  $A$  acierta 3 o más preguntas:*

$$\begin{aligned}
 P_{(3)} &= \binom{10}{3} \left(\frac{1}{5}\right)^3 \left(\frac{4}{5}\right)^7 = \frac{10!}{3!7!} \times 0.2^3 \times 0.8^7 = 0.2013 \\
 P_{(4)} &= \binom{10}{4} \left(\frac{1}{5}\right)^4 \left(\frac{4}{5}\right)^6 = \frac{10!}{4!6!} \times 0.2^4 \times 0.8^6 = 0.0881 \\
 P_{(5)} &= \binom{10}{5} \left(\frac{1}{5}\right)^5 \left(\frac{4}{5}\right)^5 = \frac{10!}{5!5!} \times 0.2^5 \times 0.8^5 = 0.0264 \\
 P_{(6)} &= \binom{10}{6} \left(\frac{1}{5}\right)^6 \left(\frac{4}{5}\right)^4 = \frac{10!}{6!4!} \times 0.2^6 \times 0.8^4 = 0.0055 \\
 P_{(7)} &= \binom{10}{7} \left(\frac{1}{5}\right)^7 \left(\frac{4}{5}\right)^3 = \frac{10!}{7!3!} \times 0.2^7 \times 0.8^3 = 0.0008 \\
 P_{(8)} &= \binom{10}{8} \left(\frac{1}{5}\right)^8 \left(\frac{4}{5}\right)^2 = \frac{10!}{8!2!} \times 0.2^8 \times 0.8^2 = 0.0001 \\
 P_{(9)} &= \binom{10}{9} \left(\frac{1}{5}\right)^9 \left(\frac{4}{5}\right)^1 = \frac{10!}{9!1!} \times 0.2^9 \times 0.8^1 = 0.0000 \\
 P_{(10)} &= \binom{10}{10} \left(\frac{1}{5}\right)^{10} \left(\frac{4}{5}\right)^0 = \frac{10!}{10!0!} \times 0.2^{10} \times 0.8^0 = 0.0000
 \end{aligned}$$

$$P_{(X \geq 3)} = P_{(3)} + P_{(4)} + P_{(5)} + P_{(6)} + P_{(7)} + P_{(8)} + P_{(9)} + P_{(10)} = 0.3222 = \alpha$$

Es decir, para el valor del parámetro en la hipótesis nula ( $p=0.20$ ) la probabilidad de acertar tres o más veces es 0.3222. Dicho de otra manera, la probabilidad de que acierte tres o más veces una persona que dice números al azar es de 0.3222. Por lo tanto la probabilidad de rechazar la hipótesis nula siendo cierta es de  $\alpha = 0.3222$  que es el nivel de significación con el que se está haciendo esta primera prueba.

*Prueba 2) Rechazar  $H_0$  si  $A$  acierta 5 o más veces:*

$$\begin{aligned}
 P_{(5)} &= \binom{10}{5} \left(\frac{1}{5}\right)^5 \left(\frac{4}{5}\right)^5 = \frac{10!}{5!5!} \times 0.2^5 \times 0.8^5 = 0.0264 \\
 P_{(6)} &= \binom{10}{6} \left(\frac{1}{5}\right)^6 \left(\frac{4}{5}\right)^4 = \frac{10!}{6!4!} \times 0.2^6 \times 0.8^4 = 0.0055
 \end{aligned}$$

$$P_{(7)} = \binom{10}{7} \left(\frac{1}{5}\right)^7 \left(\frac{4}{5}\right)^3 = \frac{10!}{7!3!} \times 0.2^7 \times 0.8^3 = 0.0008$$

$$P_{(8)} = \binom{10}{8} \left(\frac{1}{5}\right)^8 \left(\frac{4}{5}\right)^2 = \frac{10!}{8!2!} \times 0.2^8 \times 0.8^2 = 0.0001$$

$$P_{(9)} = \binom{10}{9} \left(\frac{1}{5}\right)^9 \left(\frac{4}{5}\right)^1 = \frac{10!}{9!1!} \times 0.2^9 \times 0.8^1 = 0.0000$$

$$P_{(10)} = \binom{10}{10} \left(\frac{1}{5}\right)^{10} \left(\frac{4}{5}\right)^0 = \frac{10!}{10!0!} \times 0.2^{10} \times 0.8^0 = 0.0000$$

$$P_{(X \geq 5)} = P_{(5)} + P_{(6)} + P_{(7)} + P_{(8)} + P_{(9)} + P_{(10)} = 0.0328 = \alpha$$

Es decir, para el valor del parámetro en la hipótesis nula ( $p=0.20$ ) la probabilidad de acertar cinco o más veces es 0.0328. Dicho de otra manera, la probabilidad de que acierte cinco o más veces una persona que dice números al azar es de 0.0328. Por lo tanto la probabilidad de rechazar la hipótesis nula siendo cierta es de  $\alpha = 0.0328$  que es el nivel de significación con el que se está haciendo esta segunda prueba.

*Prueba 3) Rechazar  $H_0$  si A acierta 8 o más veces:*

$$P_{(8)} = \binom{10}{8} \left(\frac{1}{5}\right)^8 \left(\frac{4}{5}\right)^2 = \frac{10!}{8!2!} \times 0.2^8 \times 0.8^2 = 0.0001$$

$$P_{(9)} = \binom{10}{9} \left(\frac{1}{5}\right)^9 \left(\frac{4}{5}\right)^1 = \frac{10!}{9!1!} \times 0.2^9 \times 0.8^1 = 0.0000$$

$$P_{(10)} = \binom{10}{10} \left(\frac{1}{5}\right)^{10} \left(\frac{4}{5}\right)^0 = \frac{10!}{10!0!} \times 0.2^{10} \times 0.8^0 = 0.0000$$

$$P_{(X \geq 8)} = P_{(8)} + P_{(9)} + P_{(10)} = 0.0001 = \alpha$$

Es decir, para el valor del parámetro en la hipótesis nula ( $p=0.20$ ) la probabilidad de acertar ocho o más veces es 0.0001. Dicho de otra manera, la probabilidad de que acierte ocho o más veces una persona que dice números al azar es de 0.0001. Por lo tanto, la probabilidad de rechazar la hipótesis nula siendo cierta es de  $\alpha=0.0001$  que es, por tanto, el nivel de significación con el que se está haciendo esta tercera prueba.

Resumiendo, se tiene que, con respecto al *Error de Tipo I*

<i>Prueba</i>	<i>Regla de decisión</i>	<i>Error Tipo I</i>
1	<i>Rechazar <math>H_0</math> si <math>X \geq 3</math></i>	$\alpha = 0.3222$
2	<i>Rechazar <math>H_0</math> si <math>X \geq 5</math></i>	$\alpha = 0.0328$
3	<i>Rechazar <math>H_0</math> si <math>X \geq 8</math></i>	$\alpha = 0.0001$

El valor máximo de  $\alpha$  es el valor de la probabilidad de obtener un resultado dentro de la región de rechazo con el valor del parámetro que indica la hipótesis nula. Por esta razón, la  $H_0$  incluirá, necesariamente, la igualdad y el nivel de significación ( $\alpha$ ) se define como la probabilidad de rechazar la hipótesis nula siendo cierto el valor del parámetro en la  $H_0$ .

Desde este punto de vista, esta claro que la prueba más conveniente sería la tercera, pues la probabilidad de rechazar la hipótesis nula siendo cierta, es muy pequeña; mientras que en la primera prueba, es muy alta la probabilidad de rechazar la hipótesis nula siendo cierta.

En términos del problema del examen de A, si se adopta la primera regla de decisión, habrá un 32% de personas que, sin tener conocimiento de la materia, se le apruebe. Mientras que con la tercera regla, solo aprobaría uno de cada 10000 personas sin conocimiento de la materia, por lo tanto, esta claro que si una persona responde correctamente ocho o más preguntas, de diez, es porque muy probablemente conozca la materia.

Más adelante se verá que esta conclusión es errónea.

Pero el comportamiento de una prueba está caracterizado por su función de potencia. Supóngase que para A,  $p = 0.6$ , es decir, que la probabilidad de que A acierte es tres veces superior a la de una persona sin conocimientos, en ese caso se tendría que para cada una de las pruebas la función de potencia correspondiente es

<i>Prueba</i>	<i>Función de potencia</i>
1	$\delta 1_{(0.6)} = P(X \geq 3 \text{ cuando el valor del parámetro es } p=0.6)$
2	$\delta 2_{(0.6)} = P(X \geq 5 \text{ cuando el valor del parámetro es } p=0.6)$
3	$\delta 3_{(0.6)} = P(X \geq 8 \text{ cuando el valor del parámetro es } p=0.6)$

En las siguientes páginas se hallan estas funciones de potencia.

*Prueba 1) Rechazar  $H_0$  si A acierta 3 o más veces, siendo  $p = 0.6$*

$$P_{(3)} = \binom{10}{3} \left(\frac{3}{5}\right)^3 \left(\frac{2}{5}\right)^7 = \frac{10!}{3!7!} \times 0.6^3 \times 0.4^7 = 0.0425$$

$$P_{(4)} = \binom{10}{4} \left(\frac{3}{5}\right)^4 \left(\frac{2}{5}\right)^6 = \frac{10!}{4!6!} \times 0.6^4 \times 0.4^6 = 0.1115$$

$$P_{(5)} = \binom{10}{5} \left(\frac{3}{5}\right)^5 \left(\frac{2}{5}\right)^5 = \frac{10!}{5!5!} \times 0.6^5 \times 0.4^5 = 0.2006$$

$$P_{(6)} = \binom{10}{6} \left(\frac{3}{5}\right)^6 \left(\frac{2}{5}\right)^4 = \frac{10!}{6!4!} \times 0.6^6 \times 0.4^4 = 0.2150$$

$$P_{(7)} = \binom{10}{7} \left(\frac{3}{5}\right)^7 \left(\frac{2}{5}\right)^3 = \frac{10!}{7!3!} \times 0.6^7 \times 0.4^3 = 0.2150$$

$$P_{(8)} = \binom{10}{8} \left(\frac{3}{5}\right)^8 \left(\frac{2}{5}\right)^2 = \frac{10!}{8!2!} \times 0.6^8 \times 0.4^2 = 0.1209$$

$$P_{(9)} = \binom{10}{9} \left(\frac{3}{5}\right)^9 \left(\frac{2}{5}\right)^1 = \frac{10!}{9!1!} \times 0.6^9 \times 0.4^1 = 0.0403$$

$$P_{(10)} = \binom{10}{10} \left(\frac{3}{5}\right)^{10} \left(\frac{2}{5}\right)^0 = \frac{10!}{10!0!} \times 0.6^{10} \times 0.4^0 = 0.0060$$

$$P_{(X \geq 3)} = P_{(3)} + P_{(4)} + P_{(5)} + P_{(6)} + P_{(7)} + P_{(8)} + P_{(9)} + P_{(10)} = 0.9877 = \alpha$$

Como  $\delta_{(0.6)}$  es

$$\delta_{(0.6)} = P_{(X \geq 3, \text{ cuando el valor del parámetro es } 0.6)} = 0.9877$$

Y por tanto  $\beta$  es

$$\beta = 1 - \delta_{(0.6)} = 1 - 0.9877 = 0.0123$$

Esto indica que, con un valor del parámetro de la  $H_1$  ( $p = 0.6$ ), la probabilidad de acertar tres o más veces es de 0.9877. Esta primera prueba tiene, por tanto, una potencia de  $\delta_{(0.6)} = 0.9877$ , por lo que la probabilidad de Error Tipo II es igual a  $1 - 0.9877 = 0.0123$ . Si la probabilidad de rechazar  $H_0$  si  $A$  acierta tres o más veces, siendo  $p = 0.6$ , es 0.9877; uno menos esta cantidad es la probabilidad de aceptar  $H_0$  si  $A$  acierta tres o más veces, siendo  $p = 0.6$ , es decir, la probabilidad de aceptar la hipótesis nula ( $p \leq 0.2$ ) cuando no es cierta ( $p = 0.6$ ) es de  $\beta = 0.0123$ .

Dicho con otras palabras; la probabilidad de obtener un resultado dentro de la región de aceptación con el valor del parámetro de la  $H_1$ , es 0.0123.

*Prueba 2) Rechazar  $H_0$  si  $A$  acierta 5 o más veces, siendo  $p = 0.6$*

$$P_{(5)} = \binom{10}{5} \left(\frac{3}{5}\right)^5 \left(\frac{2}{5}\right)^5 = \frac{10!}{5!5!} \times 0.6^5 \times 0.4^5 = 0.2006$$

$$P_{(6)} = \binom{10}{6} \left(\frac{3}{5}\right)^6 \left(\frac{2}{5}\right)^4 = \frac{10!}{6!4!} \times 0.6^6 \times 0.4^4 = 0.2150$$

$$P_{(7)} = \binom{10}{7} \left(\frac{3}{5}\right)^7 \left(\frac{2}{5}\right)^3 = \frac{10!}{7!3!} \times 0.6^7 \times 0.4^3 = 0.2150$$

$$P_{(8)} = \binom{10}{8} \left(\frac{3}{5}\right)^8 \left(\frac{2}{5}\right)^2 = \frac{10!}{8!2!} \times 0.6^8 \times 0.4^2 = 0.1209$$

$$P_{(9)} = \binom{10}{9} \left(\frac{3}{5}\right)^9 \left(\frac{2}{5}\right)^1 = \frac{10!}{9!1!} \times 0.6^9 \times 0.4^1 = 0.0403$$

$$P_{(10)} = \binom{10}{10} \left(\frac{3}{5}\right)^{10} \left(\frac{2}{5}\right)^0 = \frac{10!}{10!0!} \times 0.6^{10} \times 0.4^0 = 0.0060$$

$$P_{(X \geq 5)} = P_{(5)} + P_{(6)} + P_{(7)} + P_{(8)} + P_{(9)} + P_{(10)} = 0.8337 = \alpha$$

Como  $\delta_{(0.6)}$  es

$$\delta_{(0.6)} = P_{(X \geq 5, \text{ cuando el valor del parámetro es } 0.6)} = 0.8337$$

Y por tanto  $\beta$  es

$$\beta = 1 - \delta_{(0.6)} = 1 - 0.8337 = 0.1663$$

Esto indica que, con un valor del parámetro de la  $H_1$  ( $p=0.6$ ), la probabilidad de acertar cinco o más veces es de 0.8337. Esta primera prueba tiene, por tanto, una potencia de  $\delta_{(0.6)}=0.8337$ , por lo que la probabilidad de Error Tipo II es igual a  $1 - 0.8337=0.1663$ . Si la probabilidad de rechazar  $H_0$  si  $A$  acierta cinco o más veces, siendo  $p=0.6$ , es 0.8337; uno menos esta cantidad es la probabilidad de aceptar  $H_0$  si  $A$  acierta cinco o más veces, siendo  $p=0.6$ , es decir, la probabilidad de aceptar la hipótesis nula ( $p \leq 0.2$ ) cuando no es cierta ( $p=0.6$ ) es de  $\beta=0.1663$ .

Dicho con otras palabras; la probabilidad de obtener un resultado dentro de la región de aceptación con el valor del parámetro de la  $H_1$ , es 0.1663.

*Prueba 3) Rechazar  $H_0$  si  $A$  acierta 8 o más veces, siendo  $p=0.6$*

$$P_{(8)} = \binom{10}{8} \left(\frac{3}{5}\right)^8 \left(\frac{2}{5}\right)^2 = \frac{10!}{8!2!} \times 0.6^8 \times 0.4^2 = 0.1209$$

$$P_{(9)} = \binom{10}{9} \left(\frac{3}{5}\right)^9 \left(\frac{2}{5}\right)^1 = \frac{10!}{9!1!} \times 0.6^9 \times 0.4^1 = 0.0403$$

$$P_{(10)} = \binom{10}{10} \left(\frac{3}{5}\right)^{10} \left(\frac{2}{5}\right)^0 = \frac{10!}{10!0!} \times 0.6^{10} \times 0.4^0 = 0.0060$$

$$P_{(X \geq 8)} = P_{(8)} + P_{(9)} + P_{(10)} = 0.1672 = \alpha$$

Como  $\delta_{(0.6)}$  es

$$\delta_{(0.6)} = P_{(X \geq 8, \text{ cuando el valor del parámetro es } 0.6)} = 0.1672$$

Y por tanto  $\beta$  es

$$\beta = 1 - \delta_{(0.6)} = 1 - 0.1672 = 0.8328$$

Esto indica que, con un valor del parámetro de la  $H_1$  ( $p=0.6$ ), la probabilidad de acertar ocho o más veces es de 0.1672. Esta primera prueba tiene, por tanto, una potencia de  $\delta_{(0.6)}=0.1672$ , por lo que la probabilidad de Error Tipo II es igual a  $1-0.1672=0.8328$ . Si la probabilidad de rechazar  $H_0$  si  $A$  acierta ocho o más veces, siendo  $p=0.6$ , es 0.1672; uno menos esta cantidad es la probabilidad de aceptar  $H_0$  si  $A$  acierta ocho o más veces, siendo  $p=0.6$ , es decir, la probabilidad de aceptar la hipótesis nula ( $p \leq 0.2$ ) cuando no es cierta ( $p=0.6$ ) es de  $\beta=0.8328$ .

Dicho con otras palabras; la probabilidad de obtener un resultado dentro de la región de aceptación con el valor del parámetro de la  $H_1$ , es 0.8328.

Por tanto, en el supuesto de que  $p=0.6$ , es decir, que  $A$  tiene un conocimiento de la materia que le permite acertar el triple de veces que una persona que dice cualquier número al azar en cada experimento, se tendrá, con respecto al *Error de Tipo II*

<i>Prueba</i>	<i>Regla de decisión</i>	<i>Error Tipo II</i>
1	<i>Rechazar <math>H_0</math> si <math>X \geq 3</math></i>	$\alpha = 0.0123$
2	<i>Rechazar <math>H_0</math> si <math>X \geq 5</math></i>	$\alpha = 0.1663$
3	<i>Rechazar <math>H_0</math> si <math>X \geq 8</math></i>	$\alpha = 0.8328$

Según esto, el valor de  $\beta$  es el valor de la probabilidad de obtener un resultado dentro de la *región de aceptación* con el valor del parámetro que indica la hipótesis alternativa. Por esta razón,  $\beta$  se define como la probabilidad de aceptar la hipótesis nula siendo el autentico valor del parámetro el de la  $H_1$ , es decir, siendo falsa la hipótesis nula.

Desde este punto de vista, esta claro que la prueba más conveniente sería la primera, pues la probabilidad de aceptar la hipótesis nula siendo falsa, es muy pequeña; mientras que en la tercera prueba, es muy alta la probabilidad de aceptar la hipótesis nula siendo falsa. En términos del examen de  $A$ , si se adopta la tercera regla de decisión, habrá un 83% de personas que, conociendo la materia, se les suspendan. Mientras que con la primera regla, solo un 1% de personas con conocimientos podrían pasar desapercibidas suspenderlos. Por lo tanto, esta claro que una persona con conocimientos que acierta tres o más preguntas, de diez, es porque muy probablemente tenga conozca la materia.

Sin embargo, con respecto a  $\beta$ , se toma como mejor la prueba que, con respecto a  $\alpha$ , se tomo como la peor, y viceversa.

De modo que las pruebas con un nivel de significación alto, tienen un menor Error Tipo II, mientras que las pruebas con nivel de significación bajo tienen mayor Error Tipo II. Por lo que la conclusión es que no es posible minimizar simultáneamente los dos tipos de error. La estrategia más adecuada, dado este fenómeno, se discutirá más adelante. La discusión anterior ha sido para el valor autentico de  $p=0.6$  siendo el valor de la  $H_0$ ,  $p=0.2$ . Pero la conclusión obtenida es similar para otros valores de  $p$ , si se repiten los cálculos para tal como se han hecho antes.

Resumiendo, se tiene que en un juego de hipótesis, la hipótesis nula contendrá invariablemente la igualdad. Y la estrategia que se seguirá para elegir una prueba consistirá en fijar el máximo nivel de significación ( $\alpha$ ) que se está dispuestos a tolerar, y una vez fijado este nivel de significación elegir la prueba que minimice el Error Tipo II para todos los valores del parámetro  $\theta$ , en otras palabras, que maximice la función de potencia de la prueba para todos los valores del parámetro en la hipótesis alternativa.

En el caso de la hipótesis del ejemplo, supóngase que se ha decidido tomar como nivel de significación de la prueba el  $\alpha \leq 0.04$ . Esta restricción elimina automáticamente a las pruebas cuyas reglas de decisión serían

Rechazar  $H_0$  si  $X \geq 0$  ( $\alpha = 1.0$ )

Rechazar  $H_0$  si  $X \geq 1$  ( $\alpha = 0.8926$ )

Rechazar  $H_0$  si  $X \geq 2$  ( $\alpha = 0.6242$ )

Rechazar  $H_0$  si  $X \geq 3$  ( $\alpha = 0.3222$ )

Rechazar  $H_0$  si  $X \geq 4$  ( $\alpha = 0.1209$ )

Las pruebas que cumplen con la restricción son aquellas cuya regla de decisión es

Rechazar  $H_0$  si  $X \geq 5$  ( $\alpha = 0.0328$ )

Rechazar  $H_0$  si  $X \geq 6$  ( $\alpha = 0.0064$ )

Rechazar  $H_0$  si  $X \geq 7$  ( $\alpha = 0.0009$ )

Rechazar  $H_0$  si  $X \geq 8$  ( $\alpha = 0.0001$ )

Rechazar  $H_0$  si  $X \geq 9$  ( $\alpha = 0.0000$ )

Rechazar  $H_0$  si  $X \geq 10$  ( $\alpha = 0.0000$ )

Hay que elegir, por tanto, entre una estas seis pruebas. De acuerdo con la estrategia anteriormente estudiada, se elegirá aquella que tenga menor Error Tipo II para los diferentes valores de  $\theta$ . Si se hace un estudio como el hecho en las páginas anteriores se llega a la conclusión que la prueba cuya regla de decisión es

*Rechazar  $H_0$  si  $X \geq 5$  ( $\alpha = 0.0328$ )*

es la que maximiza la función de poder o equivalentemente minimiza las probabilidades de Error Tipo II para los diferentes valores de  $\theta$ , entre las pruebas con  $\alpha \leq 0.04$ .

Seleccionada la prueba de acuerdo con los criterios propuestos, sólo resta realizar el experimento (tomar la muestra) y tomar la decisión que corresponda. En este caso la prueba escogida es

<i>Regla de decisión, Rechazar</i>	$H_0$ si $X \geq 5$
<i>Nivel de significación,</i>	$\alpha = 0.0328$
<i>Región de rechazo,</i>	$R(X X = 5, 6, 7, 8, 9, 10)$

Esto quiere decir que una vez realizado el experimento se decidirá rechazar  $H_0$  si  $A$  acierta cinco o más veces y no se rechaza  $H_0$  si  $A$  acierta cuatro o menos veces.

Estos conceptos se completaran con los que se estudien en el epígrafe *Error de Tipo I, Error de Tipo II y tamaño de muestra*.

Hay que señalar, que esta manera de presentar una prueba de hipótesis es exclusivamente con fines didácticos. En un próximo epígrafe se expondrá el método de realización de una prueba de hipótesis. De todas maneras, adelantemos que, si aquí se ha fijado una región de rechazo y después hallado el nivel de significación, en la práctica se hará justo lo contrario.

El nivel de significación adoptado para la prueba se simboliza como

$$P > N$$

siendo  $N$  un número que indicará la probabilidad de la región de rechazo para el valor del parámetro de la hipótesis nula.

También es tradicional indicar los niveles de significación más habituales ( $\alpha = 0.05$ ,  $\alpha = 0.01$  y  $\alpha = 0.001$ ) con un asterisco, dos asteriscos y tres asteriscos, respectivamente.

Si se rechaza una hipótesis nula al nivel de significación de  $\alpha=0.05$ , se le añade al estadístico de prueba un asterisco (\*), ver cualquier ejemplo de este manual) y se concluye que el estadístico de prueba es *significativo* y por lo tanto se rechaza la hipótesis nula.

Si se rechaza una hipótesis nula al nivel de significación de  $\alpha=0.01$ , se le añade al estadístico de prueba dos asterisco (\*\*) y se concluye que el estadístico de prueba es *muy significativo* y por lo tanto se rechaza la hipótesis nula.

Si se rechaza una hipótesis nula al nivel de significación de  $\alpha=0.001$ , se le añade al estadístico de prueba tres asterisco (\*\*\*) y se concluye que el estadístico de prueba es *altamente significativo* y por lo tanto se rechaza la hipótesis nula.

En el caso de que no se rechace la hipótesis nula, se le añade al estadístico de prueba las siglas **sn** que significa *no significativo*, y por lo tanto se acepta la hipótesis nula.

Esta última simbología, por se la más ampliamente extendida es la que se sigue en este manual.

### Diferentes tipos de Hipótesis.-

Generalizando la situación estudiada en los epígrafes anteriores, si  $\Theta$  es un parámetro cualquiera y  $\Theta_0$  es una constante del conjunto de valores que puede tomar el parámetro, la hipótesis que se ha discutido ha sido



$$H_0 : \Theta \leq \Theta_0$$

$$H_1 : \Theta > \Theta_0$$

Una característica de estas hipótesis es que la hipótesis nula contiene más de un punto.

Además, debido al hecho de que para este juego de hipótesis la regla de decisión es

Rechazar  $H_0$  si el estadístico es mayor que una cierta cantidad se dice que es una hipótesis de la cola derecha.

Otra posible hipótesis, simétrica a la anterior, puede ser

$$H_0 : \Theta \geq \Theta_0$$

$$H_1 : \Theta < \Theta_0$$

En ésta, la regla de decisión es

Rechazar  $H_0$  si el estadístico es menor que una cierta cantidad se dice, en este caso, que es una hipótesis de la cola izquierda.

Y un tercer tipo de hipótesis son

$$H_0 : \Theta = \Theta_0$$

$$H_1 : \Theta \neq \Theta_0$$

Una característica de estas hipótesis es que la hipótesis nula contiene solo un punto.

En ésta, la regla de decisión es

Rechazar  $H_0$  si el estadístico es mayor o menor que una cantidad se dice que es una hipótesis de dos colas, lo que se estudiará detenidamente en el siguientes epígrafe.

### **Pruebas de hipótesis de dos colas.-**

En la práctica de trabajos científicos y técnicos, ocurre generalmente que se trata de poner a prueba una teoría sobre la que hay una opinión muy definida del valor verdadero del parámetro, por lo que la hipótesis nula tendrá un sólo punto y cualquier desviación de la muestra en un sentido u otro va en contra de la teoría propuesta.

Póngase como ejemplo el problema cuarto de los que se plantearon al comienzo del capítulo, este era el de probar la distribución de sexos en cierta especie de vertebrados. Si se realiza un experimento como este, cualquier persona opinará claramente que, conocido el proceso de la meiosis las condiciones fisiológicas y embriológicas de multitud de especies con reproducción sexual, se espera que en un parto salga macho o hembra con la misma probabilidad. Si la evidencia experimental no apoya esta afirmación, la teoría sobre el comportamiento genérico de los procesos citogenéticos, fisiológicos y embriológicos en dicha especie, no se podrá confirmar.

Para poner a prueba esta hipótesis el experimentador puede observar diez partos y anotar los resultados (se sabe que una muestra de diez es muy pequeña para un experimento, se hace con diez para que el ejemplo quepa en un espacio asequible).

Traduciendo esto al lenguaje estadístico, se tiene una población Binomial. Si se le llama éxito a obtener hembra, entonces la variable  $X$ , definida como el número de hembras de los diez partos, tiene una distribución *Binomial* con parámetro 10 y  $p=0.5$ . La hipótesis estadística sería

$$H_0 : p = 0.5$$

$$H_1 : p \neq 0.5$$

Es una hipótesis de dos colas pues está claro que la hipótesis alternativa será apoyada por la muestra tanto si se tienen *muchos éxitos* como si se tienen *muy pocos*. Es decir, se rechazará la hipótesis nula tanto si salen muchas hembras como si salen pocas hembras.

Por tanto, se tiene una región de rechazo formada por valores extremos de la cola izquierda y de la cola derecha de la distribución del estadístico bajo la hipótesis nula.

De acuerdo con lo anterior, valores de  $X$  cercanos a cero o cercanos a diez apoyarán la hipótesis alternativa y valores próximos a cinco, apoyarán la hipótesis nula. Para concretar lo que se entiende por valores extremos es necesario fijar el nivel de significación que se desea.

Supóngase que se ha decidido que el nivel de significación de la prueba sea  $\alpha \leq 0.04$ . Esto quiere decir que la probabilidad de rechazar  $H_0$  siendo el autentico valor del parámetro  $p=0.5$  debe ser menor o igual que 0.04.

Puesto que la región de rechazo estará integrada por valores extremos en ambas colas de la distribución, es fácil imaginarse varias maneras para elegir criterios de rechazo repartidos entre ambas colas, cumpliendo todos ellos con la restricción impuesta por el nivel de significación.

La regla que se adopte será aquella que, elegidos dos valores,  $c_1$  y  $c_2$ , bajo  $H_0$

$$P_{(X < c_1)} + P_{(X > c_2)} \leq \alpha = 0.04$$

Si se hallan las probabilidades de toda esta distribución, se tiene

$$P_{(0)} = \binom{10}{0} \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^{10} = \frac{10!}{0!10!} \times 0.5^0 \times 0.5^{10} = 0.00098$$

$$P_{(1)} = \binom{10}{1} \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^9 = \frac{10!}{1!9!} \times 0.5^1 \times 0.5^9 = 0.00977$$

$$P_{(2)} = \binom{10}{2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^8 = \frac{10!}{2!8!} \times 0.5^2 \times 0.5^8 = 0.04394$$

$$P_{(3)} = \binom{10}{3} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^7 = \frac{10!}{3!7!} \times 0.5^3 \times 0.5^7 = 0.11719$$

$$P_{(4)} = \binom{10}{4} \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^6 = \frac{10!}{4!6!} \times 0.5^4 \times 0.5^6 = 0.20508$$

$$P_{(5)} = \binom{10}{5} \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^5 = \frac{10!}{5!5!} \times 0.5^5 \times 0.5^5 = 0.24609$$

$$P_{(6)} = \binom{10}{6} \left(\frac{1}{2}\right)^6 \left(\frac{1}{2}\right)^4 = \frac{10!}{6!4!} \times 0.5^6 \times 0.5^4 = 0.20508$$

$$P_{(7)} = \binom{10}{7} \left(\frac{1}{2}\right)^7 \left(\frac{1}{2}\right)^3 = \frac{10!}{7!3!} \times 0.5^7 \times 0.5^3 = 0.11719$$

$$P_{(8)} = \binom{10}{8} \left(\frac{1}{2}\right)^8 \left(\frac{1}{2}\right)^2 = \frac{10!}{8!2!} \times 0.5^8 \times 0.5^2 = 0.04394$$

$$P_{(9)} = \binom{10}{9} \left(\frac{1}{2}\right)^9 \left(\frac{1}{2}\right)^1 = \frac{10!}{9!1!} \times 0.5^9 \times 0.5^1 = 0.00977$$

$$P_{(10)} = \binom{10}{10} \left(\frac{1}{2}\right)^{10} \left(\frac{1}{2}\right)^0 = \frac{10!}{10!0!} \times 0.5^{10} \times 0.5^0 = 0.00098$$

Si se adopta como regla de decisión

$$\text{Rechazar } H_0 \text{ si } X \leq 0 \text{ o si } X \geq 10$$

Se está usando un nivel de significación de

$$\alpha = P_{(0 \leq X \leq 10)} = P_{(0)} + P_{(10)} = 0.000196$$

Es decir, un  $\alpha$  muy bajo (en correspondencia  $\beta$  será muy alto). Pero esta regla no cumple con la condición que se ha impuesto al principio, esta es  $\alpha \leq 0.04$ .

Si se adopta como regla de decisión

*Rechazar  $H_0$  si  $X \leq 1$  o si  $X \geq 9$*

Se está usando un nivel de significación de

$$\alpha = P_{(1 \geq X \geq 9)} = P_{(0)} + P_{(1)} + P_{(9)} + P_{(10)} = 0.0215$$

Está más próximo del nivel de significación impuesto al principio de la experiencia, este es el de  $\alpha \leq 0.04$ .

Si se adopta como regla de decisión

*Rechazar  $H_0$  si  $X \leq 2$  o si  $X \geq 8$*

Se está usando un nivel de significación de

$$\alpha = P_{(2 \geq X \geq 8)} = P_{(0)} + P_{(1)} + P_{(2)} + P_{(8)} + P_{(9)} + P_{(10)} = 0.10938$$

Es un nivel de significación mucho mas elevado que el máximo que se ha impuesto al principio de la experiencia.

Efectuadas las observaciones anteriores, es fácil comprobar que una regla que cumple con la restricción especificada por el nivel de significación, y que hace iguales las probabilidades de que el estadístico tome un valor cualquiera de la región de rechazo de ambas colas, es la determinada por la regla de decisión

*Rechazar  $H_0$  si  $X \leq 1$  o si  $X \geq 9$*

Es decir, que  $c_1=1$  y  $c_2=9$ . Por lo que se define

$$\begin{aligned} R_1 &= (X|X \geq c_1) = R_1 = (X|X = 0.1) \\ R_2 &= (X|X \leq c_2) = R_2 = (X|X = 9.10) \end{aligned}$$

Y la región de rechazo viene dada por  $R=R_1+R_2$

Es decir, para el valor del parámetro en la hipótesis nula ( $p=0.5$ ) la probabilidad de que nazcan una o menos hembras, o nueve o más hembras es de 0.0215. Dicho de otra manera, siendo la probabilidad de hembra 0.5, la probabilidad de que nazcan una o menos hembras o nueve o mas hembras, en diez partos, es de 0.0215. Por lo tanto la probabilidad de rechazar la hipótesis nula siendo cierta es de  $\alpha = 0.0215$  que es el nivel de significación más próximo al máximo impuesto en esta prueba.

Hay que hacer constar, que no se ha conseguido el nivel de significación exacto que se impuso al principio de la experiencia por dos motivos.

- 1) Uno de ellos es que, al ser una distribución discreta, toma valores concretos (números naturales) y por ello, no ha podido haber un acercamiento mayor al nivel de significación impuesto al principio de la experiencia. Esto no ocurrirá con distribuciones continuas, como es la normal, que es el supuesto con el que se va a hacer todas las pruebas

de hipótesis excluyendo las del Capítulo de **Datos Categóricos** (Capítulo 20).

- 2) El otro motivo, el más importante, por el que no se ha conseguido llegar más próximo a la  $\alpha$  impuesta, es que el tamaño de muestra es muy pequeño ( $n=10$ ). Recuérdese, que se ha hecho con este tamaño de muestra para que el ejemplo ocupara poco espacio de este manual, pero en la realidad, un problema como este requiere un tamaño de muestra mucho mayor, en cuyo caso se hubiera echo la prueba al nivel de significación exacto. Aunque si el tamaño de muestra hubiera sido mucho mayor, la prueba de hipótesis se hubiera hecho por aproximación a la normal, como se verá en el epígrafe *Prueba de hipótesis sobre  $p$  de una distribución binomial aproximándola a una normal.*

### **Sistematización del método para realizar una prueba de hipótesis.-**

Las ideas que se han plasmado en los ejemplos anteriores pueden generalizarse para ser aplicadas a cualquier modelo probabilístico, y sobre todo al *Normal*. A continuación se expone lo que debe ser la secuencia de pasos para probar una hipótesis estadística

- 1) De acuerdo con las suposiciones que se hagan sobre el fenómeno y con la forma en que se han obtenido los datos, se elige un modelo probabilístico bajo el cual se operará.
- 2) En el modelo probabilístico elegido, se identifica el parámetro sobre el cual se desea inferir basándose en la muestra.
- 3) Se decide cuales serán las hipótesis que se van a contrastar. En la práctica, las hipótesis más útiles son

Cola derecha	Cola izquierda	Dos colas
$H_0 : \Theta \leq \Theta_0$	$H_0 : \Theta \geq \Theta_0$	$H_0 : \Theta = \Theta_0$
$H_1 : \Theta > \Theta_0$	$H_1 : \Theta < \Theta_0$	$H_1 : \Theta \neq \Theta_0$

donde  $\Theta$  es el parámetro sobre el que se quiere inferir y  $\Theta_0$  es una constante elegida por el investigador entre los elementos del conjunto de valores posibles del parámetro  $\Theta$ . Recuérdese, así mismo, que el signo de igualdad siempre va en la  $H_0$

- 4) Se elige un estadístico de prueba. Forzosamente la distribución del estadístico depende del parámetro  $\Theta$ , ya que esto es lo que permite discernir cuáles valores del estadístico son los aceptables bajo la  $H_0$  y cuáles no lo son.
- 5) De acuerdo con la distribución probabilística del estadístico de prueba, se selecciona una regla de decisión que define una región de rechazo o valores críticos, la cual está integrada por valores del estadístico que son improbables bajo la  $H_0$ . El conjunto de valores que integran la región de rechazo está limitado por el nivel de significación que se desea para la prueba. El nivel de significación se debe elegir a priori y en base a él se determinar la región de rechazo.

6) Una vez que la regla de decisión y su correspondiente región de rechazo han sido elegidas, sólo resta calcular el valor del estadístico de prueba en la muestra que se tiene y determinar si pertenece el subconjunto de su espacio muestral en la región de rechazo o a la región de no rechazo. La decisión que se toma entonces es como sigue: si el valor del estadístico está en la región de rechazo, se dice que la hipótesis nula se rechaza con un nivel de significación de  $\alpha$ . Si el valor del estadístico en la muestra está en la región de no rechazo, se dice que  $H_0$  no se rechaza al nivel de significación de  $\alpha$ . A veces, en este último caso se acostumbra a decir que se acepta la  $H_0$  al nivel de significación propuesto o con una seguridad de  $(1-\alpha)\%$ , aunque esto no es correcto. En ningún caso es correcto afirmar que la  $H_0$  es falsa sin añadir un nivel de significación, y en ningún caso se puede afirmar que es cierta, sino que no existe evidencia suficiente en la muestra tomada para rechazar la  $H_0$  con el nivel de significación.

### Prueba de hipótesis sobre la media de una distribución normal.-

Mientras que no se haga notar lo contrario, la estadística inferencial que se va a estudiar se basa en el supuesto de normalidad o de aproximación a la normalidad haciendo uso del *Teorema Central del Límite*. Por lo tanto, una vez conocidos los principios básicos sobre las pruebas de hipótesis, es conveniente estudiar de manera explícita la forma de probar hipótesis sobre los parámetros de una normal. Se empezará con pruebas de hipótesis sobre la media.

Se tiene una muestra aleatoria  $X_1, X_2, \dots, X_n$  de una variable que por sus características o estudios anteriores se le supone una distribución Normal con media  $\mu$  y varianza  $\sigma^2$ . Esto quiere decir que cada valor de la variable aleatorias es  $N(\mu, \sigma^2)$ , y son independientes. Se quiere inferir sobre el parámetro  $\mu$  de la distribución empleando para ello la muestra que se tiene. Las hipótesis de interés práctico, como se expuso en la página anterior, son

Cola derecha	Cola izquierda	Dos colas
$H_0 : \mu \leq \mu_0$	$H_0 : \mu \geq \mu_0$	$H_0 : \mu = \mu_0$
$H_1 : \mu > \mu_0$	$H_1 : \mu < \mu_0$	$H_1 : \mu \neq \mu_0$

donde  $\mu_0$  es la constante o valor paramétrico de la prueba de hipótesis (véase ejemplo).

En busca de un estadístico de prueba para esta hipótesis, es lógico pensar en  $\bar{X}$  la media muestral de las observaciones, puesto que  $\bar{X}$  es la estima muestral del parámetro  $\mu$ .

Para encontrar una regiones de rechazo razonables, recuérdese que en una muestra de una distribución normal con media  $\mu$  y varianza  $\sigma^2$

$$\bar{X} \sim N \left( \mu, \frac{\sigma^2}{n} \right)$$

y, por lo tanto, los valores de  $\bar{X}$  tendrán a estar concentrados, con probabilidad alta, alrededor del valor  $\mu$ .

Con respecto a la hipótesis de la cola derecha, está claro que valores grandes de  $\bar{X}$  apoyan la  $H_1$ . Dicho de otra manera, valores positivos de  $\bar{X} - \mu_0$  apoyan la  $H_1$ , mientras que valores negativos de  $\bar{X} - \mu_0$  apoyan la  $H_0$ . Una regla de decisión razonable sería

$$\text{Rechazar } H_0 \text{ si } \bar{X} \geq c$$

donde  $c$  es una constante que debe elegirse de manera que el nivel de significación de la prueba sea el deseado. Si ese nivel es  $\alpha$ , se debe cumplir que

$$P(\bar{X} \geq c / \mu = \mu_0) = \alpha$$

Por lo que la región de rechazo es

$$R = (\bar{X} / \bar{X} \geq c)$$

de modo que esto indica que es una prueba de la cola derecha pero no indica cual es el valor de  $c$ .

Para determinar dicho valor es necesario considerar que, si bien

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

bajo  $H_0$  se tiene que

$$\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$$

y esta es la distribución que se requiere para fijar el nivel de significación. Puesto que  $c$  debe ser tal que

$$P(\bar{X} \in R / \mu = \mu_0) = \alpha$$

el valor de  $c$  debe elegirse de acuerdo a

$$P(\bar{X} \geq c / \mu = \mu_0) = P\left(\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \geq \frac{c - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right) = \alpha$$

y, que la variable típica

$$Z_o = \frac{\bar{X} - \mu_o}{\frac{\sigma}{\sqrt{n}}}$$

tiene una distribución normal típica, se trata de encontrar  $c$  tal que

$$P \left( Z_o \geq \frac{c - \mu_o}{\frac{\sigma}{\sqrt{n}}} \right) = \alpha$$

es decir

$$\frac{c - \mu_o}{\frac{\sigma}{\sqrt{n}}} = Z_\alpha$$

y despejando  $c$  se tiene

$$c = \mu_o + \frac{\sigma}{\sqrt{n}} Z_\alpha$$

Por lo que la regla de decisión, al nivel de significación deseado es

$$\text{Rechazar } H_o \text{ si } \bar{X} \geq \mu_o + \frac{\sigma}{\sqrt{n}} Z_\alpha$$

O alternativamente podría ser

$$\text{Rechazar } H_o \text{ si } Z_o = \frac{\bar{X} - \mu_o}{\frac{\sigma}{\sqrt{n}}} \geq Z_\alpha$$

siendo

$Z_\alpha$ , el valor de  $Z$  (tabla n° 1) para el nivel de significación elegido.

$\mu_o$ , el valor máximo de  $\mu$  que impone la  $H_o$ .

$n$ , el tamaño de la muestra.

Pero para hacer uso de esta región de rechazo se encuentra, a menudo, con un inconveniente, como es que se requiere que conozcamos un parámetro como es  $\sigma$  y esto va a ocurrir muy pocas veces. Lo más común es que se tenga el valor de la estima de  $\sigma$  en la muestra, es decir, la  $S$  de la muestra y se use ésta en lugar de aquella, con lo que se tiene

$$\frac{\bar{X} - \mu_o}{\frac{S}{\sqrt{n}}}$$

que no se ajusta a una distribución normal y, por lo tanto, el valor  $Z$  no da el nivel de significación deseado. Pero se sabe que esta variable sigue una distribución  $t$  de *Student*, por tanto



$$t_o = \frac{\bar{X} - \mu_o}{\frac{S}{\sqrt{n}}} \sim t_{(n-1; \alpha)}$$

De esta manera, la regla de decisión que garantiza el nivel de significación deseado es

$$\text{Rechazar } H_o \text{ si } t_o = \frac{\bar{X} - \mu_o}{\frac{S}{\sqrt{n}}} \geq t_{(n-1; \alpha)}$$

Equivalentemente, la regla de decisión puede expresarse en función del estadístico  $\bar{X}$

$$\text{Rechazar } H_o \text{ si } \bar{X} \geq \mu_o + \frac{S}{\sqrt{n}} t_{(n-1; \alpha)}$$

siendo

$t_{(n-1; \alpha)}$ , el valor de  $t$  (tabla n° 2) para el nivel de significación elegido y para  $n-1$  *grados de libertad*.

$\mu_o$ , el valor máximo de  $\mu$  que impone la  $H_o$ .

$S$ , la estima del parámetro  $\sigma$ , en la muestra.

$n$ , el tamaño de la muestra.

La prueba estadística resultante es la conocida como prueba  $t$  de Student, mientras que la que se usaría si  $\sigma$  fuese conocida se denomina prueba  $Z$ .

La diferencia entre la distribución  $t$  y la distribución  $Z$  es grande si se tienen pocos grados de libertad, es decir, cuando la muestra es pequeña. Si se tienen más de 30 grados de libertad, prácticamente, las probabilidades son las mismas en ambas distribuciones. Es por esto por lo que la prueba  $Z$  se la conoce como una prueba para grandes muestras o poblaciones, y la prueba  $t$  es una prueba para pequeñas muestras. Aunque siempre que se tenga una muestra, la prueba estrictamente correcta es la prueba  $t$ , sea cual sea el tamaño de la muestra; el consenso entre los diferentes autores es que para 30 o más grados de libertad se puede considerar como grande la muestra y utilizar la prueba  $Z$ .

Lo dicho hasta ahora para la media es concerniente a las hipótesis de la cola derecha, es decir

$$H_o : \mu \leq \mu_o$$

$$H_1 : \mu > \mu_o$$

Para las otras dos hipótesis el razonamiento es el mismo, y las reglas de decisión serían.

Para la cola izquierda

$$H_0 : \mu \geq \mu_0$$

$$H_1 : \mu < \mu_0$$

$$\text{Rechazar } H_0 \text{ si } t_o \leq -t_{(n-1); \alpha}$$

o bien si

$$\text{Rechazar } H_0 \text{ si } \bar{X} \leq \mu_0 - \frac{S}{\sqrt{n}} t_{(n-1); \alpha}$$

Y para las dos colas

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

$$\text{Rechazar } H_0 \text{ si } t_o \leq -t_{(n-1); \alpha/2}$$

ó

$$\text{Rechazar } H_0 \text{ si } t_o \geq t_{(n-1); \alpha/2}$$

o bien

$$\text{Rechazar } H_0 \text{ si } \bar{X} \leq \mu_0 - \frac{S}{\sqrt{n}} t_{(n-1); \alpha/2}$$

ó

$$\text{Rechazar } H_0 \text{ si } \bar{X} \geq \mu_0 + \frac{S}{\sqrt{n}} t_{(n-1); \alpha/2}$$

Hay que tener cuidado con los signos, dado que tanto la distribución  $t$  como la distribución  $Z$  son simétricas y en las tablas 1 y 2 vienen sólo los valores de la cola derecha, (los valores de la cola izquierda son los mismos pero con signo menos). Si la prueba es de dos colas se puede recomendar, para evitar errores, que tanto en la  $t_o$  como en la  $Z_o$  se ponga el numerador en valor absoluto (positivo) con el fin de comparar directamente si son mayores o no que los valores críticos que nos proveen las tablas. Si la prueba es de una sola cola no puede realizarse esta reducción pues el valor de  $t_o$  o de  $Z_o$  pueden ser, en valor absoluto, mayores que los valores tabulados pero de signo diferente por lo que estarían (al contrario de la prueba de dos colas) en la región de no rechazo.

### Ejemplo.-

La propaganda de cierta marca de cereales afirma que el contenido en proteína es de 7 g por cada 100 g. Para probar si esto es cierto se toma una muestra aleatoria de 10 cajas de dicho producto, se analizan dando los siguientes valores de proteína

6, 6, 4.5, 7.5, 8, 5, 3, 3.5, 6.5, 6
--------------------------------------

Suponiendo que esta variable sigue una distribución normal, la aseveración del fabricante traducido al lenguaje estadístico es  $\mu \geq 7$ , por lo que se va a probar

$$H_0: \mu \geq 7$$

$$H_1: \mu < 7$$

En la muestra se estima

$$\bar{X} = 5.6 \quad \text{y} \quad S = 1.61245$$

por lo que la prueba es

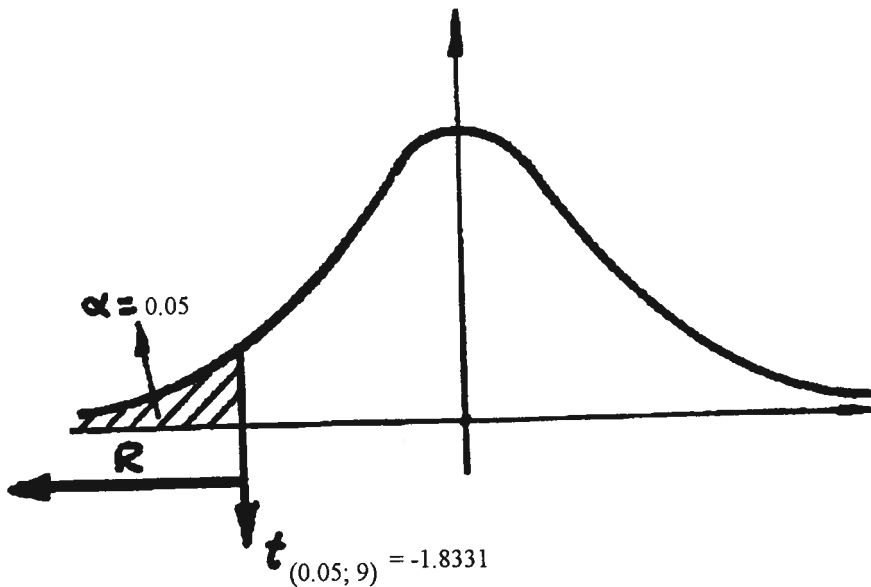
$$t_o = \frac{5.6 - 7}{\frac{1.61245}{\sqrt{10}}} = -2.7456$$

En la tabla 2 se obtiene que el valor teórico de  $t$  para una cola,  $g=9$  y  $\alpha=0.05$  es de 1.8331. Puesto que la prueba es de la cola izquierda, el valor esperado con el que hay que comparar la  $t_o$  es el valor de la tabla pero con signo negativo, es decir,  $t_{(9;0.05)} = -1.8331$ .

La región de rechazo será pues

$$\text{Rechazar } H_0 \text{ si } t_o \leq -1.8331$$

Puesto que  $t_o$  es menor que  $-1.8331$ , se rechaza la hipótesis nula, por lo que nuestro experimento contradice la afirmación de la propaganda. Gráficamente se puede ilustrar con la siguiente figura.



Este ejemplo se resolverá con el SAS en el siguiente capítulo, en el epígrafe *Generalización a más de una variable*.

### Prueba de hipótesis sobre la varianza de una distribución normal.-

Otro parámetro que aparece explícitamente en la función de distribución Normal es  $\sigma^2$ , que es la variabilidad media del fenómeno representado por el modelo. Como ejemplo, considérese dos variedades de gallinas ponedoras, las cuales tienen el mismo rendimiento ( $\mu_1 = \mu_2$ ), pero mientras una variedad tiene un rendimiento muy uniforme (distribución cercana a la media) la otra tiene un rendimiento muy variable (distribución alejada a la media). Está claro que según para lo que se vaya a utilizar ambas variedades interesará más la de mayor variabilidad o la de menor variabilidad, por lo que es deseable disponer de técnicas estadísticas para probar hipótesis sobre la varianza de una población.

Las posibles hipótesis, al igual que con la media, son:

Cola derecha	Cola izquierda	Dos colas
$H_0 : \sigma^2 \leq \sigma_0^2$	$H_0 : \sigma^2 \geq \sigma_0^2$	$H_0 : \sigma^2 = \sigma_0^2$
$H_1 : \sigma^2 > \sigma_0^2$	$H_1 : \sigma^2 < \sigma_0^2$	$H_1 : \sigma^2 \neq \sigma_0^2$

siendo  $\sigma_0^2$ , lógicamente, un número mayor de cero.

El estadístico natural para probar esta hipótesis debe estar basado en la varianza muestral  $S^2$ . De hecho, la variable aleatoria

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1; \alpha)}$$

Por lo que el estadístico que permite fijar el nivel de significación  $\alpha$  deseado es

$$\chi_o^2 = \frac{(n-1)S^2}{\sigma^2}$$

Puesto que si  $H_o$  es cierta

$$\chi_o^2 \sim \chi^2_{(n-1; \alpha)}$$

Las reglas de decisión para las diferentes hipótesis planteadas son, respectivamente

Cola derecha

$$\text{Rechazar } H_o \text{ si } \chi_o^2 \geq \chi^2_{(n-1; \alpha)}$$

Cola izquierda

$$\text{Rechazar } H_o \text{ si } \chi_o^2 \leq \chi^2_{(n-1; 1-\alpha)}$$

Dos colas

$$\text{Rechazar } H_o \text{ si } \chi_o^2 \leq \chi^2_{(n-1; 1-\alpha/2)}$$

o

$$\text{Rechazar } H_o \text{ si } \chi_o^2 \geq \chi^2_{(n-1; \alpha/2)}$$

En estas reglas se ha usado una terminología no usada hasta el momento, como es el nivel de significación  $1-\alpha$  o  $1-\alpha/2$ . La razón es que la distribución  $\chi^2$  no es simétrica y, al ser todos sus valores positivos, no tiene media cero. Esto ocasiona que se tengan que buscar directamente los puntos de la distribución en la cola izquierda. Estos valores en la cola izquierda,  $1-\alpha$  y  $1-\alpha/2$ , son abscisas tales que el área derecha de esos puntos es  $1-\alpha$  y  $1-\alpha/2$ , respectivamente. Por supuesto, las áreas de la izquierda son respectivamente  $\alpha$  y  $\alpha/2$ .

### Ejemplo.-

Se va a controlar la calidad de las mediciones del porcentaje en magnesio, de ciertos productos, de un equipo de laboratorio. De acuerdo con las normas de calidad vigentes, la máxima desviación que se debe tolerar es de 0.2% . Es decir, que si el equipo es adecuado, debe tener una  $\sigma^2 < 0.04$ .

Para establecer la variabilidad de las mediciones del equipo de laboratorio, se realizan 12 determinaciones en una solución que se sabe tiene un 5% de magnesio, con los siguientes resultados

5.1, 4.8, 5.2, 5.3, 4.7, 4.8, 4.9, 5.1, 5.2, 4.9, 4.8 y 5.3
---

La hipótesis que hay probar es

$$H_0: \sigma^2 \geq 0.04$$

$$H_1: \sigma^2 < 0.04$$

En la muestra, el valor de  $S^2 = 0.0463$ . Aunque no parece razonable que con este valor se rechace  $H_0$ , hay que realizar la prueba estadística completa. Esta es

$$\chi_o^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{11 \times 0.0463}{0.04} = 12.735$$

Usando el nivel de significación de  $\alpha = 0.05$ , se tiene que

$$\chi_{(11; 1-0.5)}^2 = \chi_{(11; 0.95)}^2 = 4.5748$$

Y puesto que esta es una prueba de la cola izquierda, la regla de decisión es

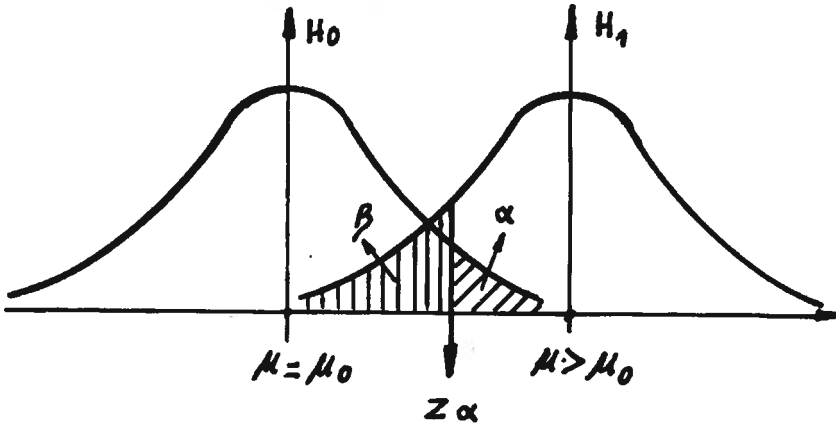
$$\text{Rechazar } H_0 \text{ si } \chi_o^2 \leq 4.5748$$

Puesto que el valor observado es  $\chi_o^2 = 12.7325$ , esta claro que no pertenece a la región de rechazo y la decisión es que, al nivel de significación de  $\alpha = 0.05$ , no se puede rechazar  $H_0$  por lo que el equipo de laboratorio no es fiable para el trabajo propuesto.

### **Error Tipo I, Error Tipo II y tamaño de muestra.-**

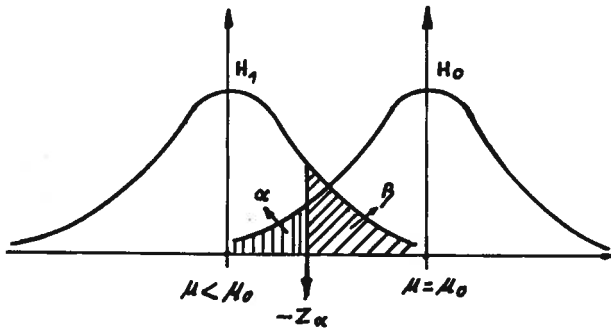
Ya se ha insinuado en varias ocasiones la interrelación entre ambos tipos de errores y el tamaño de muestra. Se va a profundizar un poco más en dichas interrelaciones a través de la distribución  $Z$  que será las más usadas, no sólo porque sea la que hay que usar por las características de los datos, sino por las aproximaciones a la Normal como consecuencia del Teorema Central del Límite, que permite aplicar una prueba  $Z$  a cualquier distribución siempre que se tenga un tamaño de muestra grande.

Si la prueba es de la cola derecha, se puede representar la prueba en la siguiente figura



Como se puede ver, mientras menor sea  $\alpha$ , mayor será el valor de  $Z_\alpha$  y, por consiguiente, mayor la probabilidad de Error Tipo II. Por el contrario, si aumenta  $\alpha$  el valor de  $Z_\alpha$  disminuye y en consecuencia  $\beta$  disminuye. Esto ya se había discutido cuando se vio que si disminuía la probabilidad de Error Tipo I aumentaba la probabilidad de Error Tipo II, y viceversa. También se ve claramente que una vez fijado el nivel de significación, la probabilidad de Error Tipo II depende del valor de  $\mu$  en el espacio de la hipótesis alternativa. Mientras más alejado esté el valor de  $\mu$  del valor de  $\mu_0$ , menor será el valor de  $\beta$ , mientras que ésta aumenta si  $\mu$  está muy cerca de  $\mu_0$ . Esto debe interpretarse de la siguiente manera: Una prueba de hipótesis que pretende distinguir entre dos valores muy cercanos, decide erróneamente una proporción mayor de veces que una que pretende distinguir entre dos valores muy alejados.

Estas mismas observaciones pueden hacerse para una prueba de la cola izquierda, como se ve en la siguiente figura



En cuanto a una prueba de dos colas, la situación es más difícil de ilustrar gráficamente, pero lo dicho anteriormente sobre los errores es igualmente válido en este caso.

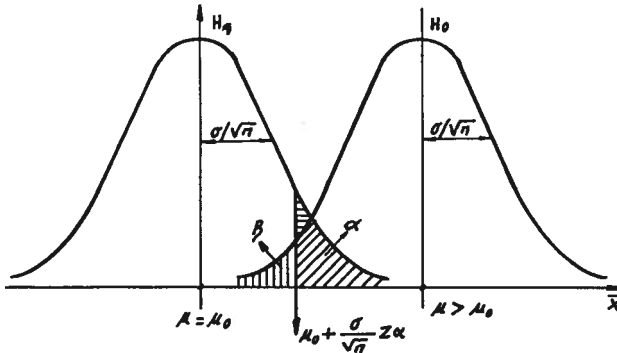
Para mostrar la influencia que tiene el tamaño de muestra sobre los errores recuérdese que en la prueba de la cola derecha la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \bar{X} \geq \mu_0 + \frac{\sigma}{\sqrt{n}} Z_\alpha$$

o lo que es lo mismo

$$\text{Rechazar } H_0 \text{ si } Z_o \geq Z_\alpha$$

Esta región de rechazo se presenta en la figura siguiente



Puesto que la distribución de  $\bar{X}$  es

$$\bar{X} \sim N \left( \mu, \frac{\sigma^2}{n} \right)$$

aumentar el tamaño de muestra tiene el efecto de disminuir la varianza de  $\bar{X}$ . El efecto de esta reducción se manifiesta en la figura anterior de dos maneras

- 1) Los valores de  $\bar{X}$  estarán más concentrados alrededor del valor de  $\mu$ , es decir, que las curvas estarían menos dispersas alrededor del valor medio.
- 2) El valor límite de la región de rechazo disminuye, ya que si el nuevo tamaño de muestra  $n$  es mayor que  $n$ , entonces

$$\mu_0 + \frac{\sigma}{\sqrt{n^*}} Z_\alpha < \mu_0 + \frac{\sigma}{\sqrt{n}} Z_\alpha$$

Esto es necesario para que el nivel de significación se mantenga, dado que las curvas son más esbeltas. El efecto combinado de estos dos hechos es disminuir el valor de  $\beta$ . La conclusión es que el valor de  $\beta$  disminuye para cualquier valor de  $\mu$  (dado un  $\alpha$ ) si se aumenta el tamaño de la muestra. La consecuencia es que aumenta  $\delta(\mu)$  en  $H_1$ , es decir, aumenta la capacidad de la prueba de rechazar hipótesis falsas.



### Ejemplo.-

Supóngase que se va a probar las hipótesis

$$H_0 : \mu \leq \mu_0$$

$$H_1 : \mu > \mu_0$$

En una situación en la que se conoce el valor de  $\alpha$ , y se está dispuesto a suponer que las condiciones necesarias para una prueba Z están justificadas. Es decir, se supone que la muestra

$$X_1, X_2, \dots, X_n$$

es de una variable aleatoria que se ajusta a una normal de manera que

$$X_i \sim N(\mu, \sigma^2)$$

La regla de decisión para la prueba es, como ya es sabido,

$$\text{Rechazar } H_0 \text{ si } \bar{X} \geq \mu_0 + \frac{\sigma}{\sqrt{n}} Z_\alpha$$

La función de poder para la prueba es entonces

$$\begin{aligned} \delta_{(\mu)} &= P(\text{Rechazar } H_0 / \mu) = P\left(\bar{X} \geq \mu_0 + \frac{\sigma}{\sqrt{n}} Z_\alpha / \mu\right) = \\ &= P\left(\bar{X} - \mu \geq \mu_0 - \mu + \frac{\sigma}{\sqrt{n}} Z_\alpha\right) = P\left(Z_0 \geq Z_\alpha + \frac{\sqrt{n}}{\sigma}(\mu_0 - \mu)\right) \end{aligned}$$

Según esta última expresión se puede hacer notar lo siguiente

a) Para todo valor de  $\mu$  en  $H_0$  ( $\mu \leq \mu_0$ ) se tiene que  $\mu_0 - \mu \geq 0$ , por lo que

$$Z_\alpha + \frac{\sqrt{n}}{\sigma} \mu_0 - \mu \geq Z_\alpha$$

esto quiere decir que para todo valor de  $\mu$  en  $H_0$

$$P\left(Z_0 \geq Z_\alpha + \frac{\sqrt{n}}{\sigma} \mu_0 - \mu\right) \leq P(Z_0 \geq Z_\alpha) = \alpha$$

Es decir, que el nivel de significación es  $\alpha$ .

b) Para todo valor de  $\mu$  en  $H_1$  ( $\mu > \mu_0$ ) se tiene que  $\mu_0 - \mu < 0$ , por lo que

$$Z_{\alpha} + \frac{\sqrt{n}}{\sigma} \mu_0 - \mu < Z_{\alpha}$$

Esto quiere decir que para todo valor de  $\mu$  en  $H_1$

$$P\left(Z_0 \geq Z_{\alpha} + \frac{\sqrt{n}}{\sigma} \mu_0 - \mu\right) > P(Z_0 \geq Z_{\alpha}) = \alpha$$

lo que implica que la probabilidad de rechazar hipótesis falsas es mayor que  $\alpha$ . Pero además implica que mientras mayor sea  $\alpha$ , mayor será  $\delta(\mu)$ , es decir, que  $\delta(\mu)$  es una función creciente de  $\mu$ .

c) También es evidente que mientras mayor sea  $\alpha$ , puesto que  $Z_{\alpha}$  es menor,  $\delta(\mu)$  aumenta y por lo tanto el Error Tipo II disminuye.

d) Finalmente, para  $\alpha$  fija, y un valor de  $\mu$  en  $H_1$  ( $\mu > \mu_0$ ), aumentar  $n$  tiene el efecto de disminuir

$$Z_{\alpha} + \frac{\sqrt{n}}{\sigma} \mu_0 - \mu$$

y, por lo tanto, de aumentar  $\delta(\mu)$  en  $H_1$ , reduciendo así la probabilidad de Error Tipo II.

### Ejemplo.-

Estas ideas se pueden ilustrar numéricamente con el siguiente ejemplo. Supongamos una población de una variable con distribución normal cuyos parámetros son  $\mu_0=5$  y  $\sigma^2=4$ . Para estudiar la relación entre las probabilidades de Error Tipo I y Error Tipo II, considérese 3 niveles de significación;  $\alpha=0.01$ ,  $\alpha=0.05$  y  $\alpha=0.10$ . El tamaño de muestra se fija en  $n=10$ . Puesto que

$$Z_{0.01} = 2.33$$

$$Z_{0.05} = 1.65$$

$$Z_{0.10} = 1.28$$

y

$$\frac{\sqrt{n}}{\sigma} = \frac{\sqrt{10}}{2} = 1.581$$

Las funciones de poder de las pruebas para los diferentes niveles de significación son, de acuerdo con la expresión que se dedujo en la página anterior

$$\delta_{0.01(\mu)} = P[Z_0 \geq 2.33 + 1.58(5 - \mu)]$$

$$\delta_{0.05(\mu)} = P[Z_0 \geq 1.65 + 1.58(5 - \mu)]$$

$$\delta_{0.10(\mu)} = P[Z_0 \geq 1.28 + 1.58(5 - \mu)]$$

Se pueden evaluar estas funciones de poder, para un amplio espectro de valores de  $\mu$ , dándole diferentes valores a la única incógnita de estas ecuaciones, es decir, a  $\mu$ .

En la siguiente tabla se presentan algunos valores

$n = 10$	$\mu$	$\delta_{0.01}(\mu)$	$\delta_{0.05}(\mu)$	$\delta_{0.10}(\mu)$
<i>Valores de <math>\mu</math> en <math>H_0</math></i>	4.4	0.0005	0.0048	0.0129
	4.7	0.0026	0.0170	0.0401
	5.0	0.0100	0.0500	0.1000
<i>Valores de <math>\mu</math> en <math>H_1</math></i>	5.3	0.0314	0.1210	0.2090
	5.6	0.0838	0.2420	0.3707
	5.9	0.1814	0.4129	0.5557
	6.2	0.3336	0.5987	0.7324
	6.5	0.5160	0.7673	0.8621
	6.8	0.6950	0.8849	0.9406
	7.1	0.8389	0.9525	0.9793
	7.4	0.9279	0.9842	0.9940
	7.7	0.9738	0.9956	0.9986
	8.0	0.9920	0.9990	0.9997

Sabiendo que  $\beta = 1 - \delta(\mu)$  se pueden expresar los valores anteriores en valores de  $\beta$

$n = 10$	$\mu$	$\beta   \alpha=0.01$	$\beta   \alpha=0.05$	$\beta   \alpha=0.10$
<i>Valores de <math>\mu</math> en <math>H_0</math></i>	4.4	0.9995	0.9952	0.9871
	4.7	0.9974	0.9830	0.9599
	5.0	0.9900	0.9500	0.9000
<i>Valores de <math>\mu</math> en <math>H_1</math></i>	5.3	0.9686	0.8790	0.7910
	5.6	0.9162	0.7580	0.6293
	5.9	0.8186	0.5871	0.4443
	6.2	0.6664	0.4042	0.2676
	6.5	0.4840	0.2327	0.1379
	6.8	0.4050	0.1151	0.0594
	7.1	0.1611	0.0475	0.0207
	7.4	0.0721	0.0158	0.0060
	7.7	0.0262	0.0044	0.0014
	8.0	0.0008	0.0010	0.0003

En estas dos tablas se constata, numéricamente, las conclusiones a las que se llegó estudiando detenidamente la expresión de la función de poder. Las más importantes, a saber

La probabilidad de rechazar hipótesis falsas es mayor que  $\alpha$

Mientras mayor sea  $\alpha$ , mayor será  $\delta(\mu)$ .

Mientras mayor sea  $\alpha$ ,  $\delta(\mu)$  aumenta.

Por lo tanto mientras mayor sea  $\alpha$ , menor es  $\beta$ .

Para analizar la última conclusión se ilustrará numéricamente el efecto del tamaño de muestra sobre el Error Tipo II. Con este fin se examina, con los mismos datos del ejemplo anterior, la función de poder de la prueba con  $\alpha=0.01$ , y cambiando el tamaño de muestra a  $n = 50$ . El dato que cambia es

$$\frac{\sqrt{n}}{\sigma} = \frac{\sqrt{50}}{2} = 3.5355 \approx 3.54$$

Por lo que la función de poder es ahora

$$\delta_{0.01(\mu)} = P[Z_0 \geq 2.33 + 3.54(5 - \mu)]$$

Dándole diferentes valores a la única incógnita de estas ecuaciones, es decir, a  $\mu$ , tal como se hizo con las anteriores funciones de poder, se pueden evaluar ésta. En la siguiente tabla se presenta esta misma función de poder para  $n = 10$ , que ya se halló en la tabla anterior, y la nueva función de poder para  $n = 50$

$\alpha = 0.01$	$\mu$	$\beta   n=10$	$\beta   n=50$
<i>Valores de <math>\mu</math> en <math>H_0</math></i>	4.4	0.9995	0.9999
	4.7	0.9974	0.9997
	5.0	0.9900	0.9900
<i>Valores de <math>\mu</math> en <math>H_1</math></i>	5.3	0.9686	0.8980
	5.6	0.9162	0.5832
	5.9	0.8186	0.1949
	6.2	0.6664	0.0274
	6.5	0.4840	0.0014
	6.8	0.4050	0.0001
	7.1	0.1611	0.0000
	7.4	0.0721	0.0000
	7.7	0.0262	0.0000
	8.0	0.0008	0.0000

Se ve claramente que el poder de la prueba se ha incrementado uniformemente para todos los valores de  $\mu$  en  $H_1$  y, consecuentemente, el Error Tipo II ha disminuido su probabilidad uniformemente.

La conclusión de este ejemplo es que, dado un nivel de significación  $\alpha$ , la forma de tener una mejor prueba, por lo que toca a la probabilidad de Error Tipo II, es aumentar el tamaño de la muestra.

### Elección del tamaño de muestra.-

En el epígrafe anterior se ha llegado a la conclusión de que, dado un  $\alpha$ , la mejor manera de tener una prueba fiable con respecto al  $\beta$  es aumentar el tamaño de la muestra, pero ¿hasta qué tamaño?

En la práctica la cuestión que se plantea es el querer saber cuál es el tamaño mínimo de muestra que se debe tomar para poder detectar una diferencia mínima verdadera

$$\delta = |\bar{X} - \mu|$$

es decir, se desea que la estima de la media no se diferencie de  $\mu$  más de  $\pm\delta$ .

$$\bar{X} = \mu \pm \delta$$

Como la curva normal va desde menos infinito a más infinito no se puede garantizar, en términos absolutos, que  $\bar{X}$  se encuentre dentro de los límites  $\mu \pm \delta$ ; sin embargo, si se puede calcular la probabilidad de que  $\bar{X}$  se encuentre dentro de estos límites, y hacer, esta probabilidad, tan grande como se quiera. En la práctica, esta probabilidad es, por lo general, del 95% o del 99%.

Si dividimos ambos miembros de la primera ecuación por el error típico de la media, se tiene

$$\frac{\delta}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

El término de la derecha de esta ecuación es el valor tipificado  $t$ , para  $gl = n-1$ , por lo que se tiene

$$\delta = \frac{\sigma}{\sqrt{n}} t_{(n-1; \alpha)}$$

de la que despejando  $n$  se tiene

$$n = \left( \frac{\sigma}{\delta} \right)^2 t_{(n-1; \alpha)}^2$$

Como se puede ver, esta ecuación requiere un conocimiento previo de  $\sigma$ , aunque la muestra no se haya obtenido aún. Este valor puede ser uno aproximado que se obtenga de trabajos anteriores con la misma o semejantes poblaciones, o bien tomando una pequeña muestra previa y utilizar  $S$  como estima de  $\sigma$ . Aunque la mayoría de las veces es suficiente con conocer el cociente  $\sigma/\delta$ .

Resumiendo, el experimentador debe decidir

Un límite máximo ( $\delta$ ) del error que quiere tolerar en la estima.

La probabilidad deseada de que la estima se encuentre dentro de este límite de error.

Un valor previo de  $\sigma$ .

En el siguiente capítulo, en el epígrafe *elección del tamaño de muestra en comparaciones de dos o más muestras*, se realiza un ejemplo

### Prueba de hipótesis de la $p$ de una distribución binomial aproximándola a la normal.-

Al comienzo del presente Capítulo se vio un par de ejemplos de prueba de hipótesis sobre la distribución Binomial haciendo uso de dicha distribución. Pero haciendo uso del teorema central del límite, esta distribución se puede aproximar a la Normal.

Aparte de la importancia que la distribución Normal tiene como modelo para interpretar muchas situaciones, su importancia se acrecienta como consecuencia del Teorema Central del Límite. Si se va a probar una hipótesis sobre la media de cualquier distribución, y se va a usar como estadístico de estima la media aritmética de la muestra, el problema fundamental, que es determinar la región de rechazo, puede abordarse usando aproximación normal, siempre que el tamaño de la muestra sea suficientemente grande.

Las posibles hipótesis son las mismas de la normal pero para el parámetro  $p$

Cola derecha	Cola izquierda	Dos colas
$H_0 : p \leq p_0$	$H_0 : p \geq p_0$	$H_0 : p = p_0$
$H_1 : p > p_0$	$H_1 : p < p_0$	$H_1 : p \neq p_0$

La estima del parámetro  $p$  es,

$$\hat{p} = \frac{X}{n}$$

por lo que este estadístico es la media de la muestra.

Será, por tanto, con esta estima,  $\hat{p}$ , con la que se determinará la región de rechazo.

Con respecto a la hipótesis de la cola derecha, está claro que valores grandes de  $\hat{p}$  apoyan la  $H_1$ . Dicho de otra manera, valores positivos de  $\hat{p} - p_0$  apoyan la  $H_1$ , mientras que valores negativos de  $\hat{p} - p_0$  apoyan la  $H_0$ . Una regla de decisión razonable sería

$$\text{Rechazar } H_0 \text{ si } \hat{p} \geq c$$

donde  $c$  es una constante que debe elegirse de manera que el nivel de significación de la prueba sea el deseado. Si ese nivel es  $\alpha$ , se debe cumplir que

$$P(\hat{p} \geq c | p = p_0) = \alpha$$

Por lo que la región de rechazo es

$$R = \{\hat{p} / \hat{p} \geq c\}$$

de modo que esto indica que es una prueba de la cola derecha pero no indica cual es el valor de  $c$ .

Para determinar dicho valor es necesario considerar que si  $n$  es grande

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

bajo  $H_0$  se tiene que

$$\hat{p} \sim N\left(p_0, \frac{p(1-p)}{n}\right)$$

y esta es la distribución que se requiere para fijar el nivel de significación. Puesto que  $c$  debe ser tal que

$$P(\hat{p} \in R \mid p = p_0) = \alpha$$

el valor de  $c$  debe elegirse de acuerdo a

$$P(\hat{p} \geq c \mid p = p_0) = P\left(\frac{\hat{p} - p_0}{\sqrt{\frac{p(1-p)}{n}}} \geq \frac{c - p_0}{\sqrt{\frac{p(1-p)}{n}}}\right) = \alpha$$

y, que la variable típica

$$Z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p(1-p)}{n}}}$$

tiene una distribución normal típica, se trata de encontrar  $c$  tal que

$$P\left(Z_0 \geq \frac{c - p_0}{\sqrt{\frac{p(1-p)}{n}}}\right) = \alpha$$

es decir

$$\frac{c - p_0}{\sqrt{\frac{p(1-p)}{n}}} = Z_\alpha$$

Despejando  $c$  se tiene

$$c = p_0 + \sqrt{\frac{p(1-p)}{n}} Z_\alpha$$

Por lo que la regla de decisión, al nivel de significación deseado es

$$\text{Rechazar } H_0 \text{ si } \hat{p} \geq p_0 + \sqrt{\frac{p(1-p)}{n}} Z_\alpha$$

O alternativamente, podría ser

$$\text{Rechazar } H_0 \text{ si } Z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p(1-p)}{n}}} \geq Z_\alpha$$

siendo

$Z_\alpha$ , el valor de  $Z$  (tabla n° 1) para el nivel de significación elegido.

$p_0$ , el valor máximo de  $p$  que impone la  $H_0$ .

$n$ , el tamaño de la muestra.

Como se puede ver, esta es una prueba  $Z$  (Normal típica) y, sin embargo, se está haciendo uso de la varianza muestral,  $S^2 = p(1-p)/n$ , y por lo tanto del error típico muestral

$$S_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

En ese caso, el hecho de que  $\sigma^2$  sea desconocida no tiene mucha importancia, puesto que si  $n$  es muy grande, como tiene que serlo,  $S$  será una estima muy próxima a  $\sigma$ , por lo que la desviación del verdadero valor será despreciable.

Lo dicho hasta ahora para la media es concerniente a las hipótesis de la cola derecha. Para las otras dos hipótesis el razonamiento es el mismo, y las reglas de decisión serían.

Para la cola izquierda

$$H_0: p \geq p_0$$

$$H_1: p < p_0$$

$$\text{Rechazar } H_0 \text{ si } Z_0 \leq -Z_\alpha$$

o bien

$$\text{Rechazar } H_0 \text{ si } \hat{p} \leq p_0 - \sqrt{\frac{p(1-p)}{n}} Z_\alpha$$

Y para las dos colas



$$H_0: p = p_0$$

$$H_1: p \neq p_0$$

Rechazar  $H_0$  si  $Z_0 \leq -Z_{\alpha/2}$

ó

Rechazar  $H_0$  si  $Z_0 \geq Z_{\alpha/2}$

o bien

$$\text{Rechazar } H_0 \text{ si } \hat{p} \leq p_0 - \sqrt{\frac{p(1-p)}{n}} Z_{\alpha/2}$$

ó

$$\text{Rechazar } H_0 \text{ si } \hat{p} \geq p_0 + \sqrt{\frac{p(1-p)}{n}} Z_{\alpha/2}$$

Como se hizo constar anteriormente, hay que tener cuidado con los signos, dado que la distribución  $Z$  es simétrica y en la tabla 1 viene solo los valores de la cola derecha, (los valores de la cola izquierda son los mismos pero con signo menos). Si la prueba es de dos colas se puede recomendar, para evitar errores, que ponga el numerador en valor absoluto (positivo) con el fin de comparar directamente si son mayores o no que los valores críticos que provee la tabla. Si la prueba es de una sola cola no puede realizarse esta reducción pues el valor de  $Z_0$  puede ser, en valor absoluto, mayor que los valores tabulados pero de signo diferente por lo que estaría (al contrario de la prueba de dos colas) en la región de no rechazo.

### Ejemplo.-

De acuerdo con el censo de 1980, la proporción de individuos menores de 40 años en cierta ciudad es del 70%. Para determinar si esa proporción se ha modificado en 2002 se toma una muestra de 800 personas, y se encuentra que de las 800 personas entrevistadas, 600 tienen menos de 40 años ¿Se puede afirmar, con un nivel de significación de  $\alpha=0.01$ , que se ha modificado la proporción de habitantes menores de 40 años?

La hipótesis que se quieren probar es

$$H_0: p = 0.7$$

$$H_1: p \neq 0.7$$

La estima de  $p$  en esta muestra es

$$\hat{p} = \frac{600}{800} = 0.75$$

Por lo que el estadístico de prueba,  $Z$ , es

$$Z_o = \frac{0.75 - 0.70}{\sqrt{\frac{0.75 \times 0.25}{800}}} = 3.266 \dots$$

Como la regla de decisión, al  $\alpha=0.01$ , es

Rechazar  $H_o$  si  $Z_o \leq -2.58$

ó

Rechazar  $H_o$  si  $Z_o \geq 2.58$

Se rechaza la hipótesis nula y se concluye que la proporción de individuos menores de 40 años se ha modificado entre 1980 y 2002, con un nivel de significación de 0.01.

Se recomienda intentar realizar esta misma prueba por la Distribución Binomial.

### **Prueba de hipótesis de la $\lambda$ de una distribución de *Poisson* aproximándola a la normal.-**

Ya se ha expresado anteriormente que la importancia de la distribución Normal no es solo como consecuencia que sirve de modelo para interpretar muchas situaciones reales, sino que su importancia se acrecienta como consecuencia del Teorema Central del Límite. Si se va a probar una hipótesis sobre la media de cualquier distribución, y se va a usar como estadístico de estima la media aritmética de la muestra, el problema fundamental, que es determinar la región de rechazo, puede abordarse usando aproximación normal, siempre que el tamaño de la muestra sea suficientemente grande.

Las posibles hipótesis son las mismas de la normal pero para el parámetro  $\lambda$

Cola derecha	Cola izquierda	Dos colas
$H_o : \lambda \leq \lambda_o$	$H_o : \lambda \geq \lambda_o$	$H_o : \lambda = \lambda_o$
$H_1 : \lambda > \lambda_o$	$H_1 : \lambda < \lambda_o$	$H_1 : \lambda \neq \lambda_o$

La estima del parámetro  $\lambda$  es

$$\lambda = n \hat{p}$$

por lo que este estadístico es la media de la muestra y, tal como se vio anteriormente, es también la estima de la varianza.

Será, por tanto, con esta estima,  $\lambda$ , con la que se determinará la región de rechazo.

Todo lo demás es lo ya deducido en el epígrafe anterior.

### Ejemplo.-

Un editor afirma que el promedio de errores por página en sus libros es menor de 0.2, es decir, que hay menos de un error por cada 5 páginas. Para probar si es cierto lo que dice el editor, se toma una muestra de 100 páginas y se cuentan el número de errores por página.

La muestra consiste, por lo tanto, en cien números

$$X_1, X_2, \dots, X_{100}$$

siendo  $X_j$  el número de errores de la página  $j$ -ésima.

El resultado de este muestreo consiste, resumidamente, en: 90 páginas sin errores, 7 páginas con un error, 2 páginas con dos errores y 1 página con tres errores.

Como se puede ver, se trata de un fenómeno que puede ser explicado por una distribución de *Poisson*. La hipótesis que se quiere probar es

$$H_0: \lambda \geq 0.2$$

$$H_1: \lambda < 0.2$$

La variable aleatoria

$$X \approx P(\lambda)$$

tiene como función de probabilidad

$$f(x_i) = \frac{e^{-\lambda} \lambda^{x_i}}{x_i!}$$

y para esta variable

$$E(X) = \lambda$$

$$\sigma^2 = \lambda$$

por lo que la hipótesis sobre  $\lambda$  es una hipótesis sobre la media de la distribución

$X_i$	$f_i$	$f_i X_i$	$f_i X_i^2$
0	90	0	0
1	7	7	7
2	2	4	8
3	1	3	9
$\Sigma$	100	14	24

La media y la varianza muestrales son

$$\bar{X} = \frac{\sum f_i X_i}{\sum f_i} = \frac{14}{100} = 0.14$$

$$S^2 = \frac{\sum f_i X_i^2 - \frac{(\sum f_i X_i)^2}{n}}{n-1} = \frac{24 - \frac{14^2}{100}}{99} = 0.22$$

El estadístico  $Z_o$  es

$$Z_o = \frac{0.14 - 0.20}{\frac{\sqrt{0.22}}{\sqrt{100}}} = -1.279$$

Si se desea probar  $H_o$  con un nivel de significación de, por ejemplo,  $\alpha=0.03$ , la prueba de la cola izquierda tiene como regla de decisión

$$\text{Rechazar } H_o \text{ si } Z_o \leq Z_{0.03}$$

De la Tabla 1 se obtiene que  $Z_{0.03} = -1.88$ , por lo que la regla queda como

$$\text{Rechazar } H_o \text{ si } -1.279 \leq -1.88$$

Como, evidentemente, no se cumple esta desigualdad, no se rechaza  $H_o$  y se concluye que no hay razón para creer que el número promedio de errores por página sea menor de 0.2, con el nivel de significación utilizado.

Si la prueba se hubiera hecho al  $\alpha = 0.05$ ,  $Z_{0.05} = 1.64$ , tampoco se cumpliría la regla de decisión, por lo que la conclusión sería la misma.

### Bibliografía:

- Dixon, J.D. y Manssey, F.J.* 1974. INTRODUCCIÓN AL ANÁLISIS ESTADÍSTICO. Ed. Del Castillo. Madrid.
- Infante Gil, S. y Zárate De Lara, G.P.* 1984. MÉTODOS ESTADÍSTICOS. Ed. TRILLAS. México.
- Gilbert, N.* 1976. ESTADÍSTICA. Ed. Interamericana. México.
- Littel, R.C., Freund, R.J. and Spector, P.C.* 1991. SAS<sup>®</sup> FOR LINEAR MODELS. SAS Institute Inc., Cary, NC, USA.
- Milton, J.S.* 1994. ESTADÍSTICA PARA BIOLOGÍA Y CIENCIAS DE LA SALUD. Ed. Interamericana-McGraw-Hill. México.
- Sokal, R.R. y Rohlf, F.J.* 1994. BIOMETRY. Ed. W.H.FREEMAN. San Francisco.
- Steel, R.* 1996. PRINCIPLES AND PROCEDURES OF STATISTICS. Ed. McGRAW-HILL Education. New York .



## **CAPÍTULO 4**

**Una/Dos poblaciones:  
*t* de Student y  $T^2$  de Hotelling**



### Una/Dos poblaciones: $t$ de Student y $T^2$ de Hotelling

Primeramente se verán las técnicas de inferencia relativas a un parámetro, utilizando un muestreo aleatorio del modelo normal. Inferir sobre los parámetros de un modelo equivale a diagnosticar sobre lo que se llama la *naturaleza* de un fenómeno.

Los datos utilizados en estas técnicas inductivas pueden ser tanto el resultado de experimentos dirigidos como el resultado de observaciones muestrales, es decir, el resultado de un *diseño experimental* o de un *diseño muestral* aleatorio simple.

Habitualmente se le denomina experimento a una situación donde la población no existe en la actualidad y el propósito de la investigación es establecer algo acerca de la población en el caso de que existiera. Una segunda situación se presenta cuando en realidad existe una población y se desea determinar mediante un muestreo aleatorio alguna característica de la población.

Un ejemplo de diseño muestral de una población puede ser: se quiere saber cual es el porcentaje medio de grasa en la leche de una determinada raza, para ello se tomará una muestra aleatoria de individuos de dicha raza y se les medirá el porcentaje de grasa, e inferir cual es el porcentaje medio de grasa en la leche de esta raza. Y un ejemplo de diseño experimental es: se quiere saber cual sería el porcentaje de grasa en la leche de cierta raza si se le alimentara con cierto concentrado, para ello se tomará una muestra aleatoria de individuos de dicha especie que serán alimentadas con dicho concentrado para medirle el porcentaje de grasa en la leche, e inferir a partir de estos resultados cual sería el porcentaje medio de grasa en la leche si existiera la población alimentada con este concentrado.

Estas dos situaciones, conceptualmente diferentes, son operativamente iguales, esto es, las técnicas estadísticas de inferencia son las mismas con la condición de que tanto los experimentos como los muestreos sean aleatorios.



Después se verán las técnicas de inferencia de los parámetros de dos poblaciones y al igual que en el caso de una población, en términos generales, se podrán tener dos poblaciones por dos motivos: porque el experimentador las ha generado mediante el uso de dos *tratamientos*, esto es, un *Diseño Experimental*; o porque ambas poblaciones o subpoblaciones existen en la estructura real de la población, por lo que se realiza la recogida de datos observacionales teniendo en cuenta esta estructura, esto es, un *Diseño Muestral*.

Al igual que para una muestra, las inferencias para estos dos tipos de diseños se realizan con las mismas técnicas analíticas, con la condición de que, en el caso del diseño experimental, los tratamientos que reciban las diferentes unidades experimentales lo sea de una manera completamente aleatoria; o, en el caso del diseño muestral, que el muestreo que se haga en cada una de las dos poblaciones sea, también, de una manera completamente aleatoria.

Pero no todos los diseño muestrales o experimentales serán completamente aleatorios (con *muestras independientes*). Habrán también diseños con *muestras emparejadas*.

En el caso de muestras emparejadas veremos que responden a un diseño experimental o muestral que incluye una *estratificación* de la población que se tiene en cuenta como consecuencia una información, y por lo tanto un análisis, más precisa que el muestreo/diseño completamente aleatorio. Para esta estratificación, la población se subdivide en varias partes o estratos de una manera no aleatoria, esto es, es el científico el que decide cuáles son los diferentes estratos con arreglo a criterios de clasificación o ambientales presentes en la población, y dentro de cada estrato se hace un diseño muestral o experimental completamente aleatorio.

Tanto si se trata de diseños experimentales como si se trata de diseños muestrales es muy importante, si se quiere realizar inferencias correctas, que se sepan ver estos estratos con objeto de introducirlos en el modelo y tenerlos en cuenta en el análisis.

También hay que aclarar que la exposición del presente manual estará enfocada más desde la perspectiva de los diseños muestrales que desde la perspectiva de los diseños experimentales, si bien, como se ha dicho anteriormente, las técnicas de análisis son las mismas para ambos tipos de diseños.

### **Ideas elementales sobre los Modelos.-**

Una práctica corriente en la Ciencia es tratar de explicar fenómenos naturales mediante *modelos*. Un modelo es una representación reducida de la realidad, es decir, de la complejidad de la realidad se modeliza la esencia de esta, la que se quiere explicar. Dicho de otro modo, un modelo es una interpretación de un hecho abstracto o inobservable expresado en términos de hechos no abstractos u observables. Por ejemplo, se puede decir que la *Tierra es como un globo* (esférica) y gira sobre un eje Norte-Sur. Que la Tierra es redonda es un hecho conjeturable y por lo tanto inferible pero no es un hecho observable (nadie ha visto la totalidad del planeta, ni siquiera los astronauta pueden verla en su totalidad) y para comprender ciertos hechos se utiliza un

modelo no abstracto y por lo tanto observable, esto es, que la Tierra es como el globo que tenemos sobre la mesa. Este es un modelo verbal que permite comprender y explicar multitud de observaciones y hechos empíricos, y permite prever resultados de experiencias aún no realizadas, como puede ser la de navegar siempre hacia el Oeste para volver al lugar de partida, lo que confirmará la veracidad (o falsedad) del modelo.

Modelos más perfeccionados dependen de afirmaciones matemáticas que permiten explicar los hechos observables (los datos), por medio de hechos no observables (adelantemos como hecho no observable, el término de *efecto de un factor*). El aspecto más importante de un modelo matemático es que posee las características intrínsecas del problema real.

### **El Modelo lineal aditivo.-**

Cuando se ha afirmado, anteriormente, que este manual esta enfocado desde la perspectiva de los diseños muestrales, se está haciendo hincapié en el hecho de que los diseños son, en principio, infinitos, y que ante un problema hay que estudiar cual es el modelo que nos provee de la información más pertinente con el mínimo de esfuerzo y dinero. La consecuencia de este modelo será la determinación del diseño (experimental o muestral) que se tiene que utilizar y el análisis adecuado para la precisión deseada.

La calidad de un análisis estadístico esta directamente relacionado con la adecuación del modelo que se asume para describir los datos. Un modelo debe representar adecuadamente la característica muestral de los datos y reflejar el problema biológico. Hay tres niveles de modelos.

- 1) Modelo *verdadero* que describa los datos perfectamente, sin variación residual o no explicada. El modelo verdadero nunca es exactamente conocido.
- 2) Un modelo *ideal* es una formulación del experimentador tan próximo al modelo verdadero como sea posible. El modelo ideal tendría que ser el modelo que se usara para el análisis, pero a menudo no se tiene suficiente información disponible como para aplicarlo en la practica.
- 3) Modelo *operacional* es una versión simplificada del modelo ideal, y es el que el experimentador usara en el análisis. A menudo no hay acuerdo entre los investigadores de cual es el mejor modelo operacional a usar.

Para elaborar un modelo operacional correcto, el experimentador debe construir un modelo ideal. Dadas las limitaciones de los datos y los recursos bajo los cuales debe funcionar el experimentador, el modelo operacional puede elaborarse simplificando apropiadamente el modelo ideal. Si se realizan demasiadas simplificaciones o se imponen demasiados supuestos, entonces puede ocurrir que el análisis no sea valido.

Siguiendo esta aproximación para construir un modelo, el experimentador debe conocer exactamente los supuestos que debe imponer, lo que implicará la mayor o menor calidad del modelo operacional. Esto quiere decir que es tan importante lo que se explicita en el modelo como lo que no se incluye.

Los modelos contendrá un grupo de *factores* que afectan aditivamente al valor de la variable. Pero un modelo no solo es la expresión matemática o ecuación, sino el conjunto de supuestos, restricciones y limitaciones implícitas.

Estos supuestos incluyen detalles que no son aparentes en la expresión matemática; por ejemplo, información de los datos o de la manera en que los datos se han tomado. ¿Fueron seleccionados aleatoriamente las unidades experimentales o fueron tomados a partir de mínimos estándar? ¿Se criaron en ambientes aleatorios o en ambientes especialmente elegidos? *etc.*

El diseño muestral (o experimental), el procedimiento estadístico y la estrategia del análisis de los datos se determinan sólo después de completado el modelo en todos sus aspectos.

El modelo lineal aditivo, por tanto, trata de explicar una observación como una suma de componentes tales como la media paramétrica más unos elementos de variación aleatorios que pueden ser asociados a diferentes efectos, fuentes de variación o factores (ver más adelante) más un *error* que incluirá tanto los *errores de medida* como los *errores de muestreo* y el *error experimental*.

Si tomamos cualquier libro de estadística veremos que vienen descritos muchos tipos de análisis de la varianza correspondientes a otros tantos diseños o modelos. Este manual se dedica a desarrollar los análisis de varianza como respuesta a diseños de muestreo, en el que la labor de diseño está en la creación del modelo lineal adecuado al problema planteado y a los datos o experiencia desarrollada. Es el mismo proceso del diseño de experimentos pero con experiencias que vienen obligadas en gran parte, como es el caso de los trabajos realizados con animales domésticos o con humanos.

Como veremos más adelante, cualquier modelo o diseño, por complejo que sea, se puede catalogar como uno de estos tres tipos

- 1) Los que incluyen sólo efectos jerárquicos.
- 2) Los que incluyen sólo efectos factoriales.
- 3) Y los que contengan efectos jerárquicos y factoriales.

Dándoseles el nombre de diseño o experimento *Jerárquico*, diseño o experimento *Factorial* y experimento *Compuesto*, respectivamente.

Un efecto está jerárquico (encajado, anidado) respecto a otro cuando está subordinado a él; por ejemplo, se tiene tres machos a prueba, para ello se toman seis hembras, dos para cada macho, en este caso el factor hembra esta jerárquico al factor macho.

Unos efectos son factoriales si tienen igual categoría entre ellos, de manera que se consideran todas las combinaciones posibles entre las diferentes clases de cada uno de los factores; por ejemplo, se aplican tres tratamientos a dos razas, probando todas las razas en cada tratamiento y los tres tratamientos en cada raza, es decir, se tendrán presentes las seis combinaciones de raza y tratamiento. Aquí los efectos de raza y tratamiento son factoriales.

### **Análisis de la Varianza (ANOVA).-**

Supóngase que se quiere medir la longitud de una pieza de madera y se hacen diez mediciones por diez personas distintas. Aunque las diez personas usen el mismo instrumento de medición, lo más seguro es que haya variación entre las diez lecturas. La precisión de una lectura dependerá de lo cuidadoso que sea el lector, de su capacidad visual, de su estado físico y anímico, *etc.* La experiencia y el resultado son los mismos si esa pieza de madera es medida diez veces por la misma persona pero sin que sepa que se trata de la misma pieza, es decir, haciéndole creer que se trata de diez piezas diferentes. La divergencia entre la auténtica medida de la pieza ( $\mu$ ) y el valor medido u observado ( $X_i$ ) es el denominado *error de medida* ( $\varepsilon_i = \mu - X_i$ ).

La experiencia y resultados serían semejantes si le pedimos a diez carpinteros que nos construyan una pieza de madera de cierta longitud, o si se le pide a un carpintero que construya, en diez tiempos diferentes, diez piezas de una misma longitud. Aun utilizando los mismos instrumentos de medida y de corte, lo más probable es que haya variación entre la longitud de las diez piezas, no solo por los errores de medida que se cometan al medirlas sino por los *errores de muestreo* que se comenten a fabricarlas.

Expliquemos el error de muestreo con otro ejemplo semejante a la fabricación de diez piezas de madera de cierta longitud. Supóngase que se miden diez personas adultas del mismo sexo y la misma población. Si se toma una de estas personas y se mide diez veces, probablemente se obtengan diez medidas diferentes, aún utilizando medidores de ultrasonidos, esto es como consecuencia del error de medida, que siempre ira adjunto a las medidas de variables cuantitativas continuas.

Las medidas de una misma persona tiene una distribución normal de la que se *espera* (en términos de esperanza matemática) tenga una media ( $\mu$ ) igual a la auténtica medida de la persona, que nos es incognoscible, y de varianza  $\sigma^2$ .

Si ahora tomamos las medidas de las diez personas, muy probablemente también serán diez medidas diferentes, no solo por los errores de medida, sino porque estas diez personas serán genéticamente diferentes (los padres de unas serán más altos que los padres de otras), habrán tenido regímenes alimenticios diferentes, se habrán criado en climas diferentes, habrán tenido enfermedades diferentes, tendrán diferentes hábitos alimenticios, deportivos, *etc.* En definitiva, las personas miden diferentes por muchas causas, unas causas pueden ser más o menos aprehensibles, mientras que otras causas son incognoscibles.

Las medidas de personas adultas del mismo sexo y población, tiene una distribución normal que se *espera* (esperanza matemática) tenga de media ( $\mu$ ) igual a la de la población y varianza  $\sigma^2$ , mayor que la varianza debida exclusivamente al error de medida de una sola persona. La divergencia entre la medida de una persona y la que se *espera* que debiera tener, por la población a la que pertenece, es debido, además del error de medida, al *error de muestreo*. El error de muestreo, por tanto, se puede definir como las diferencias en las causas que han incidido en la expresión de una misma variable en diferentes unidades experimentales o individuos. Esto quiere decir que la media o esperanza matemática de la distribución,  $\mu$ , es la altura que tendrían todas las persona de la población si todas tuvieran el mismo fondo genético, se hubieran criado en el mismo punto geográfico, en la misma familia, hubieran comido y bebido lo mismo, hubieran padecido las mismas enfermedades, hubieran realizado los mismos viajes, hubieran tenido las mismas vivencias, *etc.*, en definitiva, si todas fueran la misma persona. Como esto no es así, cada una tiene una medida.

Esta es la razón por lo que se ha estado afirmando, anteriormente que en la naturaleza existe variabilidad, es decir, que una misma variable medida en diferentes experiencias u observaciones nos da medidas diferentes. También se sabe que variables como éstas (la longitud de las piezas de madera o la altura de una persona) se ajustan a una *Normal* en la que la longitud que se quiere medir se considera como una constante desconocida ( $\mu$ ). El objetivo del experimento es concluir sobre el valor de este parámetro, y las medidas, que simbolizaremos como  $X_1, X_2, \dots, X_{10}$ , son las observaciones de una variable aleatoria que, si son independientes, se tiene una muestra aleatoria que puede usarse para inferir sobre el valor del parámetro.

Cuando se realicen las medidas u observaciones, estas pueden coincidir con el valor esperado de la longitud o discrepar de él en virtud de los errores anteriormente estudiados. Si simbolizamos los errores como  $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_{10}$ , entonces cada variable aleatoria,  $X_i$  ( $i=1, \dots, 10$ ) puede representarse como la suma de dos cantidades: la longitud esperada, que se ha simbolizado como  $\mu$ , y el error correspondiente,  $\varepsilon_i$ . Es decir, que las variables aleatorias  $X_1, \dots, X_{10}$ , pueden ser explicada por el siguiente modelo lineal

$$X_i = \mu + \varepsilon_i$$

Como el concepto esencial de este modelo es el de presentar la variable observable  $X_i$  como la suma de un parámetro  $\mu$  y una variable aleatoria no observable  $\varepsilon_i$ , dando origen a una ecuación que pretende explicar el comportamiento de la variable aleatoria  $X_i$ , es por lo que en este modelo puede presentarse también de la forma

$$\varepsilon_i = X_i - \mu$$

Esto quiere decir que el error es la diferencia entre una observación y el valor verdadero del parámetro, o lo que es lo mismo, la diferencia entre la medida realizada y la medida *esperada* (esperanza matemática).

Este modelo implica varias consecuencias que es necesario considerar.

1) Es claro que deba suponerse que las  $\varepsilon_i$  pueden tomar tanto valores positivos como negativos, puesto que suponer sólo valores positivos (o sólo negativos) equivaldría a suponer que todas las personas obtienen lecturas mayores (o menores) que la verdadera.

2) Es evidente, además, que las  $\varepsilon_i$  deben ser variables aleatorias, puesto que las  $X_i$  lo son, mientras que  $\mu$  es una constante. Por tanto, si  $\mu$  es una constante y  $\varepsilon$  una variable del tipo anteriormente descrito, se tiene, utilizando las propiedades de la media y la varianza descritas en capítulos anteriores, que

$$\begin{aligned} E(\varepsilon_i) &= 0 \\ E(X_i) &= E(\mu + \varepsilon_i) = \mu + E(\varepsilon_i) = \mu \\ \sigma^2(X_i) &= \sigma^2(\mu + \varepsilon_i) = \sigma^2(\varepsilon_i) = \sigma^2 \end{aligned}$$

Por lo que el modelo para las  $\varepsilon_i$  es Normal con media cero y varianza la de la variable medida

$$\begin{aligned} X_i &\sim N(\mu, \sigma^2) \\ \varepsilon_i &\sim N(0, \sigma^2) \end{aligned}$$

De manera que inferir sobre  $\mu$  (longitud de la pieza de madera o de las personas) equivale, como se verá en el siguiente ejemplo, a inferir sobre la media de una distribución  $N(0, \sigma^2)$  usando una muestra aleatoria de tamaño  $n$  de dicha población.

Si se le suma y resta la misma cantidad a la última expresión del modelo, su valor no cambia aunque cambie de aspecto, podemos, pues, sumarle y restarle la media muestral, quedando

$$\varepsilon_i = X_i - \mu = X_i - \bar{X} + \bar{X} - \mu$$

Como se ve, la variable  $\varepsilon_i$  es tal que su valor en las diferentes repeticiones del experimento no puede registrarse, de modo que  $\varepsilon_i$  es una variable aleatoria no observable en virtud de que el valor de  $\mu$  es desconocido y, en consecuencia, la base para la inferencia sigue siendo la variable  $X_i$  a la que se le restará el valor del parámetro que se está probando. Por lo que al hablar de  $\varepsilon_i$  en vez de  $X_i$  simplemente se está haciendo un cambio en el parámetro de localización, mediante una traslación en el eje de abscisas desde el punto  $\bar{X}$  al punto 0

Si introducimos paréntesis en el expresión anterior con objeto de clarificarla, se tiene que el modelo es

$$X_i - \mu = (X_i - \bar{X}) + (\bar{X} - \mu)$$

Esta igualdad indica que el error, o desviación total con respecto a la media poblacional, de una observación, está compuesto por la desviación con respecto a la media muestral, más la desviación entre la media muestral y la media poblacional.

### Ejemplo.-

Para continuar la explicación pongamos un ejemplo con números sencillos. Supóngase que las diez medidas de la pieza de madera o las diez alturas de las personas son

9, 10, 6, 7, 10, 11, 12, 7, 10, 8
-----------------------------------

supóngase, así mismo, que se le dijo al carpintero que estas piezas debían tener una longitud de 10. O supóngase que se quiere saber si estas diez personas pertenecen a una población de altura media  $\mu=10$ .

La media muestral de estos diez números es

$$\bar{X} = \frac{9+10+6+7+10+11+12+7+10+8}{10} = 9$$

Esto quiere decir que el carpintero ha intentado hacer diez piezas de valor 10 y sin embargo las piezas miden una media de 9. O lo que es lo mismo, se quiere saber si estas diez personas pertenecen a una población de media 10 y sin embargo miden un promedio de 9.

La pregunta que se plantea es si esta divergencia es aleatoria o es debido a un proceso direccional.

Si la divergencia es aleatoria, las diez piezas de madera no miden 10 porque, aleatoriamente, unas miden más y otras menos de 10; y como consecuencia de errores del muestreo de las diez piezas, al final la media muestral mide 9. Si se hubieran hecho otras diez piezas, la media muestral daría más, menos o igual a 10. Por tanto, en el caso de que la divergencia fuera aleatoria, la componente mayor del modelo anterior sería la del primer sumando, es decir, la debida a la diferencias entre la medida y la media de la muestra.

Si la divergencia es direccional, es porque el carpintero, por cualquier motivo, ha realizado piezas de madera de tamaño inferior a 10. O porque las personas pertenecen a una población de media inferior a 10. En este caso, la componente mayor sería la del segundo sumando, es decir, la debida a la diferencia entre la media muestral y la de la población.

Intuitivamente, se puede ver cual es la componente mayor en los datos de las diez medidas hallando estas componentes en las diez medidas, por ejemplo, para la primera pieza o persona es

$$9 - 10 = (9 - 9) + (9 - 10) = (0) + (-1) = -1$$

es decir, la componente mayor, o la única en este caso, es la debida, exclusivamente, a la diferencia entre la media de la muestra y la media de la población.

Y para la tercera pieza o persona es

$$6 - 10 = (6 - 9) + (9 - 10) = (-3) + (-1) = -4$$

es decir, en esta persona, el componente mayor es la debida a la diferencia de su medida con la media muestral.

Como se ha dicho anteriormente, la esperanza matemática de estos errores es cero, es decir, se espera que los errores positivos se compensen con los negativos, de manera que la media de los errores sea cero. Si hallamos los errores de esta diez medidas restándole a cada una la medida que se espera que tenga, es decir,  $X_i - 10$ , se tendría

-1, 0, -4, -3, 0, 1, 2, -3, 0, -2
-----------------------------------

Cuya media muestral es  $\bar{X} = -1$  que es el valor de la segunda componente del modelo para todas las medidas:  $\bar{X} - \mu = 9 - 10 = -1$ .

Si observamos atentamente esta tabla de desviaciones totales se ve que hay valores tanto positivos como negativos y la mayoría de ellos tienen un valor absoluto igual o superior al valor absoluto de la segunda componente que vale en todas -1. Por tanto, se puede, intuitivamente, concluir con cierta seguridad que la mayor componente es la primera, es decir, la diferencia entre la medida y la media de la muestra, por lo que esta divergencia es, probablemente, debida al azar.

### Ejemplo.-

Por si no queda clara esta apreciación intuitiva, contrastemos con estos otros datos que responden exactamente al mismo problema de las piezas de madera o altura de personas.

9, 9, 10, 10, 8, 9, 8, 9, 10, 8
---------------------------------

La media muestral de estos diez números es la misma que la del ejemplo anterior

$$\bar{X} = \frac{9+9+10+10+8+9+8+9+10+8}{10} = 9$$

Por lo tanto tendrá el mismo valor la segunda componente del modelo, es decir,  $\bar{X} - \mu = 9 - 10 = -1$ .



Si hallamos los errores de esta diez medidas restándole a cada una la medida que se espera que tenga, es decir,  $X_i - 10$ , se tendría

$$-1, -1, 0, 0, -2, -1, -2, -1, 0, -2$$

Si observamos estas desviación se ve, que solo hay valores negativos y que la mayoría de los valores absolutos son igual o superior al valor absoluto de la segunda componente que vale en todas -1. Se puede, intuitivamente, concluir con cierta seguridad que para estos datos la mayor componente es la segunda, es decir, la diferencia entre la media de la muestra y la de la población, por lo que las divergencias son debidas, muy probablemente, a efectos direccionales.

Concluyendo, en este ejemplo se estaría más inclinado a afirmar que esta muestra de diez individuos, no pertenece a una población de media igual a 10, mientras que en el ejemplo anterior se estaría más inclinado a afirmar que la muestra tomada puede pertenecer a una población de media igual a 10.

Como se ve, se ha llegado a unas conclusiones viendo cuál es la variabilidad total de los datos, la variabilidad de las dos componentes en que se puede dividir la variabilidad total y la relación de importancia que tienen estas dos variabilidades. En definitiva se ha realizado, de una manera intuitiva, un *Análisis de la Varianza*.

Más adelante se verá la prueba estadística para esta apreciación intuitiva y se mostrará su relación con la prueba *t* del Capítulo anterior, en el que se contrastó si una muestra pertenece a una población determinada.

### Supuestos del modelo sin factores.-

Recuérdese que, antes de ver estos ejemplos del modelo  $X_i = \mu + \varepsilon_i$ , nos quedamos en la expresión

$$X_i - \mu = (X_i - \bar{X}) + (\bar{X} - \mu)$$

Puesto que esta igualdad es cierta para todos y cada uno de los valores de  $X_i$ , y teniendo en cuenta que al sumar para  $i$  un término constante, lo que se hace es repetir dicho término tantas veces como unidades indique el valor de esta letra, se puede escribir

$$\begin{aligned} \sum_i (X_i - \mu)^2 &= \sum_i [(X_i - \bar{X}) + (\bar{X} - \mu)]^2 = \\ &= \sum_i (X_i - \bar{X})^2 + \sum_i (\bar{X} - \mu)^2 + 2 \sum_i (X_i - \bar{X})(\bar{X} - \mu) = \\ &= \sum_i (X_i - \bar{X})^2 + n (\bar{X} - \mu)^2 + 2(\bar{X} - \mu) \sum_i (X_i - \bar{X}) = \\ &= \sum_i (X_i - \bar{X})^2 + n (\bar{X} - \mu)^2 \end{aligned}$$

puesto que  $\sum_i (X_i - \bar{X}) = 0$

Esta última identidad pone de relieve el hecho importante de que la suma de cuadrados de las desviaciones respecto a un número arbitrario (distinto al valor de la media muestral) siempre es mayor (en la cantidad  $n(\bar{X}-\mu)^2$ ) que la suma de los cuadrados de las desviaciones con respecto a la media muestral.

Otra consecuencia es, si lo que se quiere probar es la hipótesis  $H_0: \mu = \mu_0$  es como hacer el contraste  $H_0: \mu - \mu_0 = 0$ . Es decir, si se tiene una muestra de tamaño  $n$  que se supone aleatoria y  $X \sim N(\mu, \sigma^2)$ , y se quiere probar la hipótesis nula

$$H_0: \mu_X = \mu_0,$$

siempre se puede definir una nueva variable aleatoria,  $Y_i = X_i - \mu_0$ , tal que, si la hipótesis nula es cierta, tiene una distribución  $Y \sim N(0, \sigma^2)$ , por lo que la variable  $Y$  puede utilizarse para probar la hipótesis

$$H_0: \mu_Y = 0,$$

lo que es una prueba equivalente a la anterior, pero con la nueva variable centrada en cero. Por lo que la expresión anterior se transforma en

$$\sum_i Y_i^2 = \sum_i (Y_i - \bar{Y})^2 + n\bar{Y}^2$$

Con las dos ecuaciones anteriores se llega a la conclusión de que la suma de cuadrados total es igual a la suma de cuadrados del error muestral más la suma de cuadrados debido a la media.

$$\begin{aligned} SC_{total} &= SC_{Error} + SC_{media} \\ \sum_i (X_i - \mu)^2 &= \sum_i (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2 \\ \sum_i Y_i^2 &= \sum_i (Y_i - \bar{Y})^2 + \frac{(\sum_i Y_i)^2}{n} \end{aligned}$$

siendo  $Y_i = X_i - \mu$ .

Puesto que estos son suma de cuadrados (numeradores de varianzas) veamos cuales son sus grados de libertad para hallar las varianzas o *Cuadrados Medios*.

El valor

$$n\bar{Y}^2 = \frac{(\sum_i Y_i)^2}{n}$$

tiene un solo grado de libertad, ya que para cualquier valor dado de  $n$ , su valor sólo depende de  $\bar{Y}$  o de  $\sum_i Y_i$ .

El valor  $\sum_i Y_i$  tiene  $n$  grados de libertad, por lo que los grados de libertad de

$$\sum_i Y_i^2 - n\bar{Y}^2 = \sum_i (Y_i - \bar{Y})^2 = \sum_i e_i^2$$

son  $n-1$ .

Para contrastar si el cuadrado medio de la media es mayor que el cuadrado medio del error, se dividen ambos cuadrados medios y se tiene una prueba  $F$  de contraste de varianzas.

$$F_o = \frac{n\bar{Y}^2}{S^2}$$

puesto que

$$S^2 = \frac{\sum Y_i^2 - n\bar{Y}^2}{n-1}$$

Esta razón tendrá una distribución  $F_{(1, n-1)}$ , por lo que la regla de decisión puede ser rechazar  $H_o$  si  $F_o > F_{(1, n-1)}$ . Como el valor esperado del  $CM_\mu$  es mayor que el del  $CM_{Error}$ , la cola adecuada de la distribución será la cola derecha.

Observando detenidamente la anterior  $F_o$  se puede comprobar que es la misma expresión que la  $t$  de *Student*, de manera que

$$F_o = \frac{n\bar{Y}^2}{S^2} = \frac{\bar{Y}^2}{\frac{s^2}{n}} = t_o^2$$

pues, recuérdese que

$$t_o = \frac{\bar{Y}}{\frac{s}{\sqrt{n}}}$$

Es decir, que esta  $F_o$  no es sino el cuadrado de la expresión de la  $t_o$  de *Student*. Y dado que el cuadrado de la distribución  $t$  con  $n-1$  grados de libertad es la distribución  $F_{(1, n-1)}$ , se concluye que se trata de la misma prueba.

Resumiendo, se tiene que el análisis completo de este modelo es

<i>FV</i>	<i>gl</i>	<i>SC</i>	<i>CM</i>	<i>F</i>	<i>ECM</i>
<i>Media</i>	1	$n\bar{Y}^2$	$n\bar{Y}^2$	$\frac{n\bar{Y}^2}{S^2}$	$\sigma^2 + n\mu^2$
<i>Error</i>	$n-1$	$\sum Y_i^2 - n\bar{Y}^2$	$\frac{\sum Y_i^2 - n\bar{Y}^2}{n-1}$		$\sigma^2$
<i>Total</i>	$n$	$\sum Y_i^2$			

**Ejemplo.-**

Resolvamos ahora el segundo ejemplo propuesto en el epígrafe *análisis de la varianza (ANOVA)*. Se trataba, por ejemplo, de la altura de 10 individuos y se quiere determinar si pertenecen a una población de media paramétrica ( $\mu = 10$ )

9, 9, 10, 10, 8, 9, 8, 9, 10, 8

$$\bar{Y} = \frac{-1-1+0+0-2-1-2-1+0-2}{10} = -1$$

$$\sum Y_i^2 = (-1)^2 + (-1)^2 + \dots + (0)^2 + (-2)^2 = 16$$

FV	gl	SC	CM	F	ECM
Media	1	$10(-1)^2=10$	10	$\frac{10}{0.66667} = 15.00^{**}$	$\sigma^2 + \eta\mu^2$
Error	10-1	16-10=6	$\frac{6}{9} = 0.667$		$\sigma^2$
Total	10	10			

Las F de contraste son  $F_{(1,10; 0.05)} = 4.96$  y  $F_{(1,10; 0.01)} = 10.0$ .

Al ser la  $F_0$  significativa, indica que esta muestra no pertenece a una población de  $\mu = 10$ . Por lo que se ha concluido lo que se dedujo intuitivamente en el epígrafe *análisis de la varianza (ANOVA)*.

**Archivo del programa SAS (C4-1.SAS).-**

```

title 'Análisis de la varianza sin factores';
options ls=75 ps=60;
data thote;
infile 'c4-1.dat';
input indi altura;
haltura=altura-10;
A = 1;
proc means mean stderr t prt;
var haltura;
run;
proc glm;
class A;
model haltura = A / noint;
run;
    
```

**Archivo de datos (C4-1.DAT).-**

1 9  
2 9

3	10
4	10
5	8
6	9
7	8
8	9
9	10
10	8

Se han creado dos nuevas variables, **HALTURA** y **A**.

La variable *altura* es la medida de la altura menos el valor paramétrico para la que se esta probando, en este caso 10. Si el valor medio de la muestra es 10, se espera que la media de las *altura* sea cero.

La variable **A=1** define una variable que se usará después para asignar todos los valores a una misma clase o factor **CLASS A**.

Se ha introducido el procedimiento **MEANS**, que hace la prueba *t* de Student para la media de una muestra igual a cero, y comparar con la salida del procedimiento **GLM**.

La opción **NOINT** del modelo en el procedimiento **GLM** indica que la prueba debe hacerse para la media *hprotein* igual a cero.

En el archivo de resultados (**C4-1.LST**), en la salida del procedimiento *means*, observamos la media, el error típico, el valor de *t* y su nivel de significación:

Mean	Std Error	T	Prob> T
-1.0000000	0.2581989	-3.8729833	0.0038

$t = -3.87298$  que es significativo al  $P < 0.01$  y se concluye lo que ya se había visto a simple vista, esto es, que esta muestra no pertenece a una población de media 10.

Y con respecto a la salida del procedimiento *glm*, observamos una salida semejante a la tabla de la resolución manual del ejemplo

Dependent Variable: HALTURA						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	1	10.0000000	10.0000000	15.00	0.0038	
Error	9	6.0000000	0.6666667			
Uncorrected Total	10	16.0000000				
R-Square						
0.000000	C.V.	-81.64966	Root MSE	0.81650	HALTURA Mean	
					-1.00000	
Source	DF	Type III SS	Mean Square	F Value	Pr > F	
1	1	10.0000000	10.0000000	15.00	0.0038	

$F = 15.00$ , que es significativo al  $P < 0.01$  y se concluye lo que ya se había visto a simple vista, esto es, que esta muestra no pertenece a una población de media 10.

Se puede observar que estos dos procedimientos son redundantes, ya que

$$F = t^2 = (-3.87298)^2 = 15.00$$

y, lógicamente, ambos tienen exactamente el mismo nivel de significación.

### Ejemplo.-

Si se utiliza este mismo programa y se cambia el archivo de datos por el siguiente

#### Archivo de datos (C4-1.DAT)

1	9.
2	10
3	6
4	7
5	10
6	11
7	12
8	7
9	10
10	8

que se corresponde con el primer ejemplo del epígrafe *análisis de la varianza (ANOVA)*, se observa que los valores de  $t$  y de  $F$  son, respectivamente,  $-1.627ns$  y  $2.65ns$ , con lo que se comprueba que era correcta la apreciación intuitiva de que no hay evidencia suficiente para concluir que esta muestra no pertenece a una población de media paramétrica 10

### Generalización a más de una variable.-

La generalización del análisis anterior a más de una variable no tiene complicación conceptual alguna si se utilizan paquetes estadísticos. Esto es así porque la única complicación que tiene el uso de más de una variable es metodológica, esto es, se requiere de cálculos más complicados, procedentes del álgebra lineal que los paquetes estadísticos realizan automáticamente y el usuario no aprecia la complicación, siendo los conceptos los mismos que se han expuesto anteriormente.

En el caso de varias variables se realizará una prueba para contrastar los valores observados en un grupo de variables de una muestra de sujetos con los parámetros teóricos de la población conocidos.

La prueba clásica es la **Lambda de Wilks** que puede probarse por medio de una  $F$  de **Snedecor**. Existen otros estadísticos, el SAS provee también la **traza de Pillai**, de la **traza de Hotelling-Lawley** y de la **mayor raíz de Roy**. El valor de  $F$  de estos cuatro estadísticos acostumbra a dar el mismo valor, por lo que las conclusiones son siempre las mismas sea cual sea el estadístico utilizado.

### La $T^2$ de Hotelling.-

La  $T^2$  de Hotelling puede considerarse como una generalización de la  $t$  de Student para el caso de más de una variable. Se conoce como  $T^2$  de Hotelling de una muestra al contraste que se establece entre los valores observados en un grupo de variables de una muestra de sujetos con unos parámetros teóricos de la población conocidos

La  $T^2$  se obtiene del valor  $F$  de la lambda de Wilks mediante la transformación

$$T^2 = F \frac{V(N-1)}{N-V}$$

siendo  $V$  el número de variables.

### Ejemplo.-

La propaganda de cierta marca de cereales afirma que, por cada 100 g, contiene 7 g de proteínas, 80.1 g de hidratos de carbono, 4.6 g de grasas, 3.4 g de fibra y 0.2 g de sodio.

Para probar si esto es cierto se toma una muestra aleatoria de 10 cajas de dicho producto, se analizan dando los siguientes valores

<i>Proteínas</i>	<i>Hidratos C</i>	<i>Grasas</i>	<i>Fibra</i>	<i>Sodio</i>
6.0	70.5	3.9	3.1	0.30
6.0	69.9	4.5	3.2	0.40
4.5	80.9	4.2	4.0	0.02
7.5	85.3	3.9	2.9	0.10
8.0	79.4	5.1	3.0	0.00
5.0	78.0	4.3	3.6	0.05
3.0	80.6	4.7	3.9	0.10
3.5	80.7	5.0	2.9	0.25
6.5	78.0	4.1	3.0	0.30
6.0	82.0	4.5	3.7	0.23

## Archivo del programa SAS (C4-2.SAS).-

```
title 'Análisis de la varianza sin factores, multivariante';
options ls=75 ps=60;
data thote;
infile 'c4-2.dat';
input indi protei hcarbo grasa fibra sodio;
hprotei=protei-7;
hcarbo=hcarbo-80.1;
hgrasa=grasa-4.6;
hfibra=fibra-3.4;
hsodio=sodio-0.2;
a = 1;
proc means mean stderr t prt;
var hprotei hcarbo hgrasa hfibra hsodio;
run;
proc glm;
class A;
model hprotei hcarbo hgrasa hfibra hsodio = a / noint;
manova h = A;
run;
```

## Archivo de datos (C4-2.DAT).-

1	6.0	70.5	3.9	3.1	0.30
2	6.0	69.9	4.5	3.2	0.40
3	4.5	80.9	4.2	4.0	0.02
4	7.5	85.3	3.9	2.9	0.10
5	8.0	79.4	5.1	3.0	0.00
6	5.0	78.0	4.3	3.6	0.05
7	3.0	80.6	4.7	3.9	0.10
8	3.5	80.7	5.0	2.9	0.25
9	6.5	78.0	4.1	3.0	0.30
10	6.0	82.0	4.5	3.7	0.23

Como en el ejemplo univariante, se han creado nuevas variables, las se denominan como las variables originales pero precedidas por una **h** y la **A**.

La variable precedidas por una **h** contienen la cantidad de los compuestos respectivos menos el valor paramétrico para el que se esta probando, 7,80.1, 4.6, 3.4 y 0.2, respectivamente. Si el valor medio de la muestra se corresponden con dichos valores paramétricos, se espera que las medias valgan cero.

La variable **A=1** define una variable que se usará después para asignar todos los valores a una misma clase o factor **CLASS A**.

Se ha introducido el procedimiento **MEANS**, que hace la prueba *t* de *Student* para las medias de una muestra igual a cero, y comparar con la salida del procedimiento **GLM**.

La opción **NOINT** del modelo en el procedimiento **GLM** indica que la prueba debe hacerse para las medias igual a cero.



El estamento **MANOVA** indica que haga el análisis multivariante con todas las variables especificadas en el modelo, y la opción **H=A** indica que haga la prueba para A.

En el archivo de resultados (C4-2.LST) observamos que la primera salida es la del procedimiento **MEANS** donde nos da los valores de la *t* de *Student* y su significación para cada variable, tal como se vio en el ejemplo anterior.

Como se ve el valor de la proteína es estadísticamente diferentes del que afirma la publicidad (este ejemplo se resolvió a mano en el capítulo anterior, en el epígrafe *Prueba de hipótesis sobre la media de una distribución normal*), mientras que los valores de los demás compuestos si se corresponden con la publicidad.

Las siguiente salidas son las del procedimiento **GLM** y son los análisis de la varianza de cada una de las variables por separado, tal como se vio en el ejemplo anterior.

Se puede comprobar que los valores de las **F** de estos análisis son igual al cuadrado de las *t* de los análisis anteriores y, por tanto, las significaciones son exactamente las mismas.

El final de la salida del procedimiento **glm** es el análisis multivariante para las cinco variables analizadas conjuntamente. Este análisis se realiza por los cuatro estadísticos citados anteriormente.

Manova Test Criteria and Exact F Statistics for the Hypothesis of no Overall A Effect						
H = Type III SS&CP Matrix for A    E = Error SS&CP Matrix						
	S=1	M=1.5	N=1.5			
Statistic	Value	F	Num DF	Den DF	Pr > F	
Wilks. Lambda	0.20920627	3.7800	5	5	0.0854	
Pillai.s Trace	0.79079373	3.7800	5	5	0.0854	
Hotelling-Lawley Trace	3.77997150	3.7800	5	5	0.0854	
Roy.s Greatest Root	3.77997150	3.7800	5	5	0.0854	

Como se ve, el valor de las **F** de los cuatro estadísticos es el mismo y esto ocurrirá casi siempre. Cuando el valor de las **F** sean diferentes, el valor a tomar para calcular la **T<sup>2</sup>** de *Hotelling* será la de la Lambda de *Wilks*, que siempre es un valor exacto.

El valor de la **T<sup>2</sup>** de *Hotelling* es

$$T^2 = F \frac{V(N-1)}{N-V} = 3.78 \frac{5(10-1)}{10-5} = 34.02$$

que tiene la misma significación que la **F**, por lo que es no significativa al 0.05 (aunque por los pelos, es la proteína la que más se desvía), y se concluye que, en conjunto, es cierto lo que dice la propaganda de esta marca.

## Dos Poblaciones.-

En términos generales, se puede tener dos poblaciones por dos motivos

- 1) Porque el experimentador las ha generado mediante el uso de dos *tratamientos*. Esto es, un *Diseño Experimental*.
- 2) Porque ambas poblaciones o subpoblaciones existen en la estructura real de la población, por lo que se realiza la recogida de datos observacionales teniendo en cuenta esta estructura. Esto es, un *Diseño Muestral*.

Un tratamiento puede ser una sustancia, una técnica, un proceso de fabricación, etc. La idea es que el experimentador aplica los tratamientos a unidades, que reciben el nombre de *unidades experimentales*, en cada una de las cuales se observa un *efecto* o respuesta de la variable medida al tratamiento. La aplicación de dos tratamientos, cada uno a un grupo de unidades experimentales, genera dos muestras aleatorias de poblaciones (posiblemente) distintas. El investigador desea comparar *parámetros* de las dos poblaciones así generadas con el propósito práctico de establecer la superioridad o no de un tratamiento sobre el otro.

Pero no siempre las respuestas se producen por la aplicación directa de tratamientos por parte del experimentador. En algunos casos, estas respuestas son directamente observables en estado natural, o dicho de otra manera, los *tratamientos* no son aplicados por el experimentador, sino que son situaciones reales que son muestreadas u observadas aleatoriamente, esto es, un diseño muestral. Este sería el caso de un estudio en que se quiere comparar las producciones de dos ganaderías. En este caso, los tratamientos (dos ganaderías) están determinados por una serie de condiciones que no dependen del experimentador, pero que, el investigador, ha de saber apreciar para realizar el análisis adecuado. Otros ejemplos podrían ser: los efectos de dos hospitales, de dos plantas de un hospital, de dos regiones, de dos estaciones climáticas, de dos sexos, etc.

Con arreglo al tipo de muestreo o de diseño experimental, estos tipos de estudios pueden realizarse de dos maneras: Con *muestras independientes* o con *muestras emparejadas*.

Se tiene muestras independientes, cuando se desea comparar dos poblaciones y se han tomado muestras de cada una de las poblaciones, aleatoria e independientemente. Si lo que se van a comparar son dos tratamientos, estos han tenido que ser asignados *completamente al azar* a cada unidad experimental (*Diseño completamente aleatorio*)

Se tienen muestras emparejadas o comparaciones emparejadas cuando se seleccionan pares de individuos o de cosas similares. Los tratamientos no se aplican completamente al azar, sino que lo que se aleatoriza es, dentro de cada par, cuál unidad experimental recibirá uno u otro tratamiento, pues previamente se ha decidido cuales son las parejas y que un tratamiento se aplica a un miembro del par y el otro tratamiento al segundo miembro del par. Los miembros de un par pueden ser dos

individuos del mismo sexo y de la misma camada, dos esquejes de la misma planta, dos costados del mismo individuo, en definitiva, dos unidades experimentales uniformes o idénticos para la variable que se va a estudiar (*Diseño de bloques aleatorios*). La aplicación más clara de datos apareados es el *autoemparejamiento*, en el que una misma unidad experimental es medida dos veces. Por ejemplo, se podrían tomar, al azar,  $n$  individuos y medirle la presión sanguínea antes y después de tomar un *vasodilatador*.

### Comparación de las varianzas de dos poblaciones normales.-

Dado que para comparar las medias (parámetro de posición) de dos poblaciones normales se necesita un contraste previo de varianzas (parámetro de dispersión) que indique si las dos distribuciones tienen la misma forma, es por lo que se empieza por esta prueba.

A continuación se discutirán procedimientos para comparar las varianzas de dos poblaciones Normales. Hay que subrayar el hecho de que los métodos que se van a estudiar dependen fuertemente de la suposición de que las poblaciones de las que se toman las muestras son Normales. Es decir, que las inferencias no son robustas con respecto a la suposición de la distribución.

La estructura de los datos necesarios para la comparación de dos varianzas se presenta de la siguiente manera

- a)  $X_{11}, X_{12}, \dots, X_{1n_1}$  es una muestra aleatoria de  $N(\mu_{(X_1)}, \sigma_{(X_1)}^2)$
- b)  $X_{21}, X_{22}, \dots, X_{2n_2}$  es una muestra aleatoria de  $N(\mu_{(X_2)}, \sigma_{(X_2)}^2)$
- c) Las dos muestras son independientes

Los estimadores naturales para  $\sigma_{(X_1)}^2$  y  $\sigma_{(X_2)}^2$  son

$$S_{(X_1)}^2 = \frac{\sum_i (X_{1i} - \bar{X}_1)^2}{n_1 - 1}$$

$$S_{(X_2)}^2 = \frac{\sum_i (X_{2i} - \bar{X}_2)^2}{n_2 - 1}$$

Y además se sabe que

$$\frac{(n_1 - 1) S_{(X_1)}^2}{\sigma_{(X_1)}^2} \sim \chi_{(n_1 - 1; \alpha)}^2$$

$$\frac{(n_2 - 1) S_{(X_2)}^2}{\sigma_{(X_2)}^2} \sim \chi_{(n_2 - 1; \alpha)}^2$$

En consecuencia, dado que las dos muestras son independientes

$$\frac{\frac{S^2_{(X_1)}}{\sigma^2_{(X_1)}}}{\frac{S^2_{(X_2)}}{\sigma^2_{(X_2)}}} \sim F_{\left(\frac{n_1-1}{n_2-1}; \alpha\right)}$$

y

$$\frac{\frac{S^2_{(X_2)}}{\sigma^2_{(X_2)}}}{\frac{S^2_{(X_1)}}{\sigma^2_{(X_1)}}} \sim F_{\left(\frac{n_2-1}{n_1-1}; \alpha\right)}$$

De aquí que la distribución  $F$  permita probar hipótesis sobre los tamaños relativos de las dos varianzas. Las hipótesis de interés son

$$\text{a) } H_0: \frac{\sigma^2_{(X_1)}}{\sigma^2_{(X_2)}} \leq 1$$

$$H_1: \frac{\sigma^2_{(X_1)}}{\sigma^2_{(X_2)}} > 1$$

$$\text{b) } H_0: \frac{\sigma^2_{(X_1)}}{\sigma^2_{(X_2)}} \geq 1$$

$$H_1: \frac{\sigma^2_{(X_1)}}{\sigma^2_{(X_2)}} < 1$$

$$\text{c) } H_0: \frac{\sigma^2_{(X_1)}}{\sigma^2_{(X_2)}} = 1$$

$$H_1: \frac{\sigma^2_{(X_1)}}{\sigma^2_{(X_2)}} \neq 1$$

Que son equivalentes a las hipótesis

$$\text{a) } H_0: \sigma^2_{(X_1)} \leq \sigma^2_{(X_2)}$$

$$H_1: \sigma^2_{(X_1)} > \sigma^2_{(X_2)}$$

$$\text{b) } H_0: \sigma^2_{(X_1)} \geq \sigma^2_{(X_2)}$$

$$H_1: \sigma^2_{(X_1)} < \sigma^2_{(X_2)}$$

$$\text{c) } H_0: \sigma^2_{(X_1)} = \sigma^2_{(X_2)}$$

$$H_1: \sigma^2_{(X_1)} \neq \sigma^2_{(X_2)}$$

De acuerdo con esto se tiene que

$$F_o = \frac{\sigma_{(X_1)}^2}{\sigma_{(X_2)}^2}$$

y

$$F_o^* = \frac{\sigma_{(X_2)}^2}{\sigma_{(X_1)}^2}$$

tienen ambas la distribución  $F$ , la primera con *gl numerador* =  $n_1-1$  y *gl denominador* =  $n_2-1$  y la segunda con *gl numerador* =  $n_2-1$  y *gl denominador* =  $n_1-1$ . Es decir, que los *gl* del numerador y del denominador se intercambian. Por lo tanto, una prueba con nivel de significación  $\alpha$ , en las hipótesis a) puede hacerse mediante el estadístico  $F_o$ , con la regla de decisión

$$\text{Rechazar } H_0 \text{ si } F_o \geq F_{(n_1-1, n_2-1; \alpha)}$$

con lo que se tiene una prueba de la cola derecha.

Para el juego de hipótesis b) podría utilizarse también el estadístico  $F_o$ , dando como resultado una prueba de la cola izquierda. Sin embargo, como ya se sabe, lo usual es que las tablas den valores críticos de la cola derecha de la distribución. Por esta razón el juego de hipótesis b) se prueba mediante el estadístico  $F_o^*$ , con la regla de decisión de

$$\text{Rechazar } H_0 \text{ si } F_o^* \geq F_{(n_2-1, n_1-1; \alpha)}$$

es decir, se coloca en el numerador de la  $F_o$  y de la  $F_o^*$  la varianza de mayor valor para que la prueba sea siempre de la cola derecha.

Pero en ciertos problemas de pruebas de homogeneidad de dos varianzas, puede no haber razón para suponer cuál varianza será mayor, así no hay razón para calcular  $F_o$ , en vez de  $F_o^*$ . Este es el caso c), que al ser una prueba de dos colas, tiene en la hipótesis alternativa la desigualdad de varianzas. Esta es la prueba que se tendrá que realizar, previamente, en los siguientes apartados para saber qué error típico de diferencias de medias se tiene que usar para la prueba  $t$ .

Si la hipótesis planteada, por tanto, es la c), el criterio de prueba  $F$  es

$$F_o = \frac{S_{mayor}^2}{S_{menor}^2}$$

Y la regla de decisión será, por tanto

$$\text{Rechazar } H_0 \text{ si } F_o \geq F_{(n_1-1, n_2-1; \alpha/2)}$$

siendo  $n_1-1$  los grados de libertad correspondientes a la muestra que presenta una mayor varianza y  $n_2-1$  los grados de libertad de la muestra que presenta una menor varianza.

Como se ve, en el criterio de rechazo, los niveles de significación con que se mira en la tabla  $F$  (Tabla 3) se tienen que multiplicar por dos, de manera que si el  $F_o$

es mayor que el  $F_{(\alpha=0.025)}$ , el nivel de significación con el que realmente se está trabajando es del  $\alpha=0.05$ ; si es mayor que  $F_{(\alpha=0.05)}$ , el nivel de significación es  $\alpha=0.1$ , y así sucesivamente.

### Ejemplo.-

En los siguientes epígrafes se tendrá que hacer pruebas de homogeneidad de varianzas, por lo que se verán los ejemplos en ese momento.

### Comparación de las medias de dos poblaciones normales.-

La mayoría de las veces, la investigación se diseña para descubrir y evaluar *diferencias entre efectos de los tratamientos*, más que para medir los efectos por sí mismos. Lo que se quiere saber, por ejemplo, es la superioridad o no, en el grado de mejoría a cierta enfermedad, que propicia un fármaco sobre otro; o la superioridad, en el aumento de peso, que produce una dieta sobre otra; o la mayor o menor presión sanguínea de individuos consumidores o no de cierto producto alimenticio, etc.

Al aplicar (o al muestrear) los efectos de dos tratamientos sobre una misma variable, se generan dos poblaciones, o distribuciones, normales, con la misma forma (misma varianza) y, probablemente, situadas en puntos diferentes (diferentes medias). Si se quiere conocer la diferencia real entre los efectos de estos dos tratamientos, se tendría que conocer las dos poblaciones y comprobar si se solapan completamente (no existe diferencia en los efectos) o, por el contrario, una está desplazada a la derecha de la otra en el eje de las abscisas (el efecto de un tratamiento es mayor que el efecto del otro tratamiento). Como no es posible conocer las poblaciones al ser infinitas, se tendrá que tomar muestras y a partir de ellas inferir (*estimar*) sobre el parámetro de posición de ambas poblaciones, es decir, sobre la diferencia de las medias.

Para comparar  $\mu_1$  y  $\mu_2$ , se toma una muestra de tamaño  $n_1$  de la primera población y una muestra de tamaño  $n_2$  de la segunda población, teniendo, por tanto

$$X_{11}, X_{12}, \dots, X_{1n_1} \sim N(\mu_1, \sigma^2)$$

$$X_{21}, X_{22}, \dots, X_{2n_2} \sim N(\mu_2, \sigma^2)$$

Se desea saber si las medias de las dos poblaciones son iguales, es decir, si  $\mu_1 = \mu_2$

Sean los dos parámetros de cada población,  $\mu_1$ ,  $\mu_2$  y  $\sigma^2$ , las medias y las varianzas de dichas poblaciones. Para comparar  $\mu_1$  y  $\mu_2$ , hay que hacerlo a través de sus medias muestrales o estimas respectivas,  $\bar{X}_1$  y  $\bar{X}_2$ .

Se puede probar si las medias de ambas poblaciones son iguales, contrastando

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

para ello se utiliza las diferencias de las estimas  $\bar{X}_1 - \bar{X}_2$ .

Esta claro que un valor de  $\bar{X}_1 - \bar{X}_2$  cercano a cero apoyará la hipótesis nula (medias iguales), mientras que un valor alejado del cero apoyará la hipótesis alternativa (medias diferentes).

Recuérdese que la distribución de la media muestral es

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

siendo la desviación típica de dicha distribución

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

el denominado *error típico* de la media.

Basándose en los mismos principios, se desprende fácilmente que la distribución muestral de  $\bar{X}_1 - \bar{X}_2$  es normal con media

$$\begin{aligned}\mu_{(\bar{X}_1 - \bar{X}_2)} &= \frac{\sum_i (X_{1i} - X_{2i})}{n} = \frac{\sum_i X_{1i} - \sum_i X_{2i}}{n} = \\ &= \frac{\sum_i X_{1i}}{n} - \frac{\sum_i X_{2i}}{n} = \mu_1 - \mu_2\end{aligned}$$

y con varianza

$$\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}$$

Por lo que

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right)$$

siendo

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$$

el *error típico de la diferencia de medias*.

Por tanto, los valores de  $\bar{X}_1 - \bar{X}_2$  tenderán a estar concentrados, con probabilidad alta, alrededor del valor  $\mu_1 - \mu_2$ .

De esta manera, la distribución muestral del estadístico

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{X}_1 - \bar{X}_2}}$$

es una normal típica, es decir, media cero y varianza uno. Por lo que una regla de decisión razonable para probar las hipótesis planteadas sería

$$\text{Rechazar } H_0 \text{ si } Z_o = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}} \geq Z_{(\alpha/2)}$$

siendo  $Z_\alpha$  el valor de  $Z$  (tabla 1) para el nivel de significación elegido, habitualmente para  $\alpha=0.05$ ,  $\alpha=0.01$  y  $\alpha=0.001$ .

Pero para hacer uso de esta prueba hay que conocer el valor del parámetro  $\sigma^2$ , situación poco frecuente. Lo habitual es que se desconozca el valor de la varianza y que, por tanto, haya que usar su estima a través del estadístico muestral.

Cuando se tiene una muestra no hay duda, la estima de  $\sigma^2$  es  $S^2$ . Pero cuál es la estima de  $\sigma^2$  cuando se tiene dos muestras. La respuesta es simple, la estima de  $\sigma^2$  es la media ponderada de las dos varianzas muestrales, lo que se denomina *varianza conjunta*

$$S_{1,2}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$$

Como se ve, esta expresión es la adición de la suma de cuadrados de las dos muestras dividida por la suma de los grados de libertad de ambas muestras.

Por tanto, se usará como error típico de la diferencia de medias la estima siguiente

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{S_{1,2}^2}{n_1} + \frac{S_{1,2}^2}{n_2}}$$

Quedando, ahora, la distribución muestral del estadístico como sigue

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_{\bar{X}_1 - \bar{X}_2}}$$



Obsérvese que una vez más  $t$  mide la distancia de una variable aleatoria o una media hipotética en unidades de desviación típica de la variable aleatoria, por lo que su distribución es como la  $t$  de *Student*.

Por lo que, en este caso, la regla de decisión razonable para probar las hipótesis planteadas sería

$$\text{Rechazar } H_0 \text{ si } t_o = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}} \geq t_{(n_1-1+n_2-1; \alpha/2)}$$

siendo  $t_{(n_1-1+n_2-1; \alpha/2)}$  el valor de  $t$  para  $gl = n_1-1+n_2-1$  y el nivel de significación elegido, habitualmente para  $\alpha=0.05$ ,  $\alpha=0.01$  y  $\alpha=0.001$ .

### **Supuestos paramétricos y limitaciones de la prueba $t$ -**

Hay dos supuestos o condiciones paramétricas que se han utilizado para deducir esta última expresión o estadístico de prueba  $t$  y dos limitaciones.

El primer supuesto paramétrico que se planteó para deducir esta expresión es

#### ***Las dos poblaciones son normales***

Si no se cumple este supuesto hay que utilizar alguna de las pruebas no paramétricas. Aunque hay que hacer notar que esta prueba es robusta al supuesto de normalidad, en especial en las cercanías de los valores críticos usualmente utilizados, es decir, para  $\alpha \leq 0.05$ .

El segundo supuesto paramétrico que se planteó es

#### ***Las dos poblaciones tienen varianza común, $\sigma^2$***

Si no se cumple este supuesto o condición, hay que aplicar otra prueba  $t$  que se verá más adelante, en el epígrafe *Comparación de dos medias de muestras independientes y varianzas desiguales*.

La primera limitación es que esta prueba está diseñada para dos medias y, por lo tanto, no es adecuada para generalizar a más de dos medias, ya que no existe una forma clara de definir el numerador de  $t$  para la generalización. Cabría la posibilidad, si se desea probar la diferencia entre seis medias, de hacer una prueba  $t$  de cada par, se tendrían, entonces, 15 pruebas  $t$ . Hay tres razones por las que este no es un buen procedimiento. Primeramente, porque si se usa un  $\alpha=0.05$  para contrastar varias medias por pares, el nivel de significación será mucho mayor del 0.05, puesto que, aunque todas las muestras proceden de la misma población, por término medio, el 5% de los valores de  $t$  excederán el valor crítico. Puede demostrarse que la probabilidad de que uno o más de los 15 valores independientes de  $t$  excedan a  $t_{(0.05)}$  es 40%, por lo que, si la hipótesis anterior de las seis medias iguales, se rechaza cuando algunos de los 15 valores de  $t$  exceden a  $t_{(0.05)}$ , está claro que se está usando el nivel de significación de  $\alpha=0.4$ . La segunda razón por la que no es conveniente este tipo de

prueba para más de dos poblaciones, es que hay una pérdida de precisión al estimar la varianza si se usan solamente aquellas medidas de los dos grupos que se están comparando. Esto se mejora usando una varianza combinada, Y el tercer motivo por el que no debe usarse esta prueba para comparar más de dos medias, es que para ello existe el *análisis de la varianza* que se verá más adelante.

Por lo tanto, la solución a esta limitación es bien simple. Si se tiene mas de dos poblaciones normales, se pueden comparar por medio del análisis de varianza.

La segunda limitación es que esta prueba es para dos muestra independientes como consecuencia de que el denominador de este estadístico de prueba, la  $t_o$ , es el error típico de la diferencia de medias que implica la estima de la varianza conjunta,  $S^2_{1,2}$ , de ambas muestras mediante la suma de las dos varianzas. Por lo que es conveniente comprobar si la varianza de una suma es igual que la suma de las varianzas. No es muy complicado de comprobar, tal como se ve a continuación

$$\begin{aligned}
 \sigma^2_{(X_1+X_2)} &= \frac{\sum_i [(X_{1i} + X_{2i}) - \overline{(X_1 + X_2)}]^2}{n} = \\
 &= \frac{\sum_i [(X_{1i} + X_{2i}) - \frac{\sum (X_1 + X_2)}{n}]^2}{n} = \\
 &= \frac{\sum_i [(X_{1i} + X_{2i}) - \frac{\sum X_1}{n} - \frac{\sum X_2}{n}]^2}{n} = \\
 &= \frac{\sum_i [(X_{1i} + X_{2i}) - \bar{X}_1 - \bar{X}_2]^2}{n} = \\
 &= \frac{\sum_i [(X_{1i} - \bar{X}_1) + (X_{2i} + \bar{X}_2)]^2}{n} = \\
 &= \frac{\sum_i \{ (X_{1i} - \bar{X}_1)^2 + (X_{2i} - \bar{X}_2)^2 + 2[(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)] \}}{n} = \\
 &= \frac{\sum_i (X_{1i} - \bar{X}_1)^2}{n} + \frac{\sum_i (X_{2i} - \bar{X}_2)^2}{n} + \frac{2 \sum_i (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{n} = \\
 &= \sigma^2_{X_1} + \sigma^2_{X_2} + 2 \sigma_{X_1 X_2}
 \end{aligned}$$

siendo  $\sigma_{X_1 X_2}$  la covarianza paramétrica, es decir, lo equivalente a la varianza pero para dos variable. Este parámetro mide el grado de variación o el grado de asociación en la variación de ambas variables. Su estima es

$$Cov = \frac{\sum X_1 X_2 - \frac{\sum X_1 \sum X_2}{n}}{n-1}$$

como se ve, si se sustituye  $X_2$  por  $X_1$ , la covarianza daría la varianza de  $X_1$ .

La expresión para los estadísticos muestrales de la varianza de la suma de dos variables es

$$S_{X_1+X_2}^2 = S_{X_1}^2 + S_{X_2}^2 + 2 Cov$$

Esto indica que si se trabaja con una variable compuesta de la suma de otras dos variables, la varianza de esta nueva variable será la suma de las varianzas de las variables que la componen, más un nuevo término añadido que es función del grado en que estén asociados los valores de éstas dos variables. Este valor añadido a la suma de las dos varianzas es el doble de la covarianza.

Por lo que el error típico, para la prueba  $t$  vista en el epígrafe anterior, es demasiado grande, al estar basado en la suma de las varianzas de las dos variables sin sustraerle ningún término debido a la covarianza.

Cuando las dos variables que se suman no están correlacionadas (no están asociadas, no varían conjuntamente), este término añadido de la covarianza será también cero y la varianza de la suma será simplemente la suma de las varianzas de las dos variables. Esta es la razón por la que se insiste en el supuesto de la *independencia* entre las dos variables a fin de poder sumar sus varianzas en el denominador del estadístico de prueba. De otro modo habría que haber tenido en consideración un término de covarianza que hasta ahora no se ha tenido en cuenta. Sin embargo, si se tiene datos emparejados se espera encontrar correlación entre las variables, dado que cada pareja comparte una experiencia común.

Por lo tanto, hay que tener en cuenta la información sobre el apareamiento con objeto de eliminar una fuente de variación, extraña a la finalidad de la experiencia, como es la variación existente entre pares.

### Ejemplo.-

Se está estudiando la cantidad media de leche que producen cabras de dos razas ( $G$  y  $M$ ). Se quiere saber si existe diferencia en la producción de leche entre las dos razas. Los datos son

$X_1$  ( $G$ ): 515, 564, 647, 538, 573, 556, 528, 522  
 $X_2$  ( $M$ ): 657, 598, 643, 634, 595, 637, 643

$$\begin{array}{ll} n_1 = 8 & n_2 = 7 \\ \bar{X}_1 = 555.3750 & \bar{X}_2 = 629.5714 \\ S_1^2 = 1796.5535 & S_2^2 = 563.2857 \end{array}$$

Siendo razas diferentes, podría ocurrir que la forma de las distribuciones también fueran diferentes, por lo que es conveniente realizar el contraste previo de varianzas

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

$$F_o = \frac{1796.5535}{563.2857} = 3.189ns$$

$$F_{(7,6; 0.05/2)} = 5.70$$

no se rechaza la hipótesis nula, por lo que, efectivamente, las dos varianzas son iguales.

Se pasa, entonces, a contrastar lo que indicará si la producción de leche es la misma o no en las dos razas, es decir, la prueba  $t$

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

$$S_{1,2}^2 = \frac{7 \times 1796.5535 + 6 \times 563.2857}{13} = 1227.353$$

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{1227.353 \left( \frac{1}{8} + \frac{1}{7} \right)} = 18.1316$$

$$t_o = \frac{555.375 - 629.5714}{18.1316} = -4.0921^{**}$$

$$t_{(13; 0.01)} = 3.0123$$

Por tanto se rechaza la  $H_0$  con un nivel de significación de  $\alpha=0.01$ . Por lo que se puede afirmar que la cantidad de leche producida es diferente en estas dos razas.

#### Archivo del programa SAS (C4-3.SAS) .-

```

title 'prueba t, muestras independientes';
options ls=75 ps=60;
data pruebat;
infile 'c4-3.dat';
input raza $ leche @@;
proc ttest;
  class raza;
  var leche;
run;

```

#### Archivo de datos (C4-3.DAT).-

```

G 515 M 657
G 564 M 598
G 647 M 643
G 538 M 634
G 573 M 595
G 556 M 637

```

G 528 M 643  
G 522

En el archivo de resultados, C4-3.LST, la primera tabla muestra el tamaño, la media, la desviación típica, el error típico y el valor máximo y mínimo de cada muestra.

TTEST PROCEDURE						
Variable: LECHE						
RAZA	N	Mean	Std Dev	Std Error	Minimum	Maximum
G	8	555.3750000	42.38577086	14.98563300	515.0000000	647.0000000
M	7	629.5714286	23.73364098	8.97047311	595.0000000	657.0000000

La siguiente tabla muestra el valor de  $t$ , sus grados de libertad y significación para el caso de que las varianzas de las dos muestras sean desiguales y para el caso en que sean iguales.

Variances	T	DF	Prob> T
Unequal	-4.2482	11.2	0.0013
Equal	-4.0921	13.0	0.0013

Y en la última línea del archivo de resultados hace la prueba de hipótesis para contrastar si son iguales o diferentes las varianzas de las dos muestras

For H0: Variances are equal, F. = 3.19 DF = (7,6) Prob>F. = 0.1789

la  $F = 3.19$  es no significativa al  $\alpha = 0.05$ , por lo que se pueden considerar iguales las varianzas de las dos muestras.

El valor de  $t$  para las varianzas iguales vale -4.092 y es significativo al  $P < 0.0013$  (menor que 0.05 y 0.01), por lo que se concluye que es diferente la cantidad de leche que producen las dos razas.

### Dos poblaciones con análisis de varianza.-

En el caso del análisis de varianza de un factor con sólo dos niveles se puede demostrar que el estadístico de prueba  $F$  es igual al cuadrado del estadístico de prueba  $t$  para el caso en que se compara dos medias de muestras con la misma varianza.

### Ejemplo.-

Si se resuelve por medio del análisis de varianza el ejemplo anterior se tiene

## Archivo del programa SAS (C4-4.SAS).-

```
title 'Análisis de la varianza, 1 vía, dos tratamientos';
options ps=65 ls=75;
data prt_f;
infile 'c4-3.dat';
input raza $ leche @@;
title .Análisis de varianza para resolución del mismo
problema.;
proc anova;
  class raza;
  model leche = raza;
run;
```

El archivo de resultados muestra que el valor de la  $F$  es de 16.75

Source	DF	Anova SS	Mean Square	F Value	Pr > F
RAZA	1	20552.4107	20552.4107	16.75	0.0013

que efectivamente

$$t = \sqrt{16.75} = 4.092$$

y es significativo al mismo nivel de  $P < 0.0013$ , por lo que se concluye lo mismo que con la prueba  $t$ .

## Generalización multivariante.-

En el caso de más de una variable, la  $T^2$  de *Hotteling* se puede calcular de dos maneras

a)

$$T^2 = (N - 2) \lambda \left( \frac{1}{\lambda} - 1 \right)$$

b)

$$T^2 = \left( \frac{(n_1 + n_2 - 2)V}{n_1 + n_2 - V - 1} \right)$$

siendo  $\lambda$  la Lambda de Wilks y  $V$  el número de variables.

### Ejemplo.-

Se ha medido la longitud del tórax y la longitud del élitro de dos especies de chinche de las plantas, *Halticus oleracea* y *Halticus carduorum*. Los datos son

<i>Halticus oleracea</i>		<i>Halticus carduorum</i>	
tórax	élitros	tórax	élitros
180	245	181	305
192	260	158	237
217	276	184	300
221	299	171	273
171	239	181	297
192	262	181	308
213	278	177	301
192	255	198	308
170	244	180	286
201	276	177	299
		176	317
		192	312
		164	265

### Archivo del programa SAS (C4-5.SAS).-

```
title 'Análisis de la varianza, 1 vía, 2 tratamientos, multivariante';
options ls=75 ps=60;
data pru_t;
infile 'c4-5.dat';
input especie $ torax elitro;
proc ttest;
class especie;
var torax elitro;
run;
proc glm;
class especie;
model torax elitro=especie;
manova h=especie;
run;
```

### Archivo de datos (C4-5.DAT).-

```
oleracea 180 245
oleracea 192 260
oleracea 217 276
oleracea 221 299
oleracea 171 239
oleracea 192 262
oleracea 213 278
oleracea 192 255
oleracea 170 244
oleracea 201 276
carduorum 181 305
carduorum 158 237
carduorum 184 300
carduorum 171 273
```

```

carduorun 181 297
carduorun 181 308
carduorun 177 301
carduorun 198 308
carduorun 180 286
carduorun 177 299
carduorun 176 317
carduorun 192 312
carduorun 164 265

```

Primero se ha puesto el procedimiento **TTEST** que hará la prueba *t* de *Student* para comparar las medias del tórax y del élitro de las dos especies.

Las salidas del procedimiento **ttest** y las primeras salidas del procedimiento **glm** son como las explicadas en el ejemplo anterior.

La dos primeras salidas del archivo de resultados, **C4-5.LST**, son las de las pruebas *t* para sendas variables, en donde se observa que para la variable *torax*,  $t = 2.736^*$  y para la variable *elitra*,  $t = 3.343^{**}$ . Esto es, las dos variables son diferentes en ambas especies.

Las siguientes salidas son las del procedimiento GLM en las que se puede observar que los valores de la **F** para los análisis univariantes son el cuadrado de las *t* del procedimiento TTEST para varianzas iguales.

La última tabla del archivo de resultados es la del análisis multivariante, esta es

Manova Test Criteria and Exact F Statistics for the Hypothesis of no Overall ESPECIE Effect						
H = Type III SS&CP Matrix for ESPECIE    E = Error SS&CP Matrix						
	S=1	M=0	N=9			
Statistic	Value	F	Num DF	Den DF	Pr > F	
Wilks. Lambda	0.16458329	50.7595	2	20	0.0001	
Pillai.s Trace	0.83541671	50.7595	2	20	0.0001	
Hotelling-Lawley Trace	5.07595110	50.7595	2	20	0.0001	
Roy.s Greatest Root	5.07595110	50.7595	2	20	0.0001	

La  $T^2$  de *Hotelling* vale

$$T^2 = 21 \left( \frac{1}{0.16458} - 1 \right) = 106.5952^{***}$$

o bien

$$T^2 = 50.7595 \left( \frac{(10+13-2)2}{10+13-2-1} \right) = 106.5949^{***}$$

que es significativa al 0.001 por lo que se concluyen que ambas medidas, consideradas en conjunto, son diferentes.



## Comparación de medias muestrales usando observaciones emparejadas.-

Las unidades experimentales se aparean a menudo. Por ejemplo, dos raciones alimenticias pueden compararse utilizando dos animales de cada una de 10 camadas de cerdos, asignando al azar la ración que recibe cada animal dentro de cada una de las diez camadas; o puede compararse la cantidad de aceite en la semillas de girasol en dos variedades cultivadas en parcelas juntas de 12 localidades diferentes; o puede medirse la presión sanguínea en individuos antes y después de suministrado un supuesto vasopresor. El emparejamiento se ha hecho antes de comenzar el experimento basándose en que las respuestas serían similares si no hubieran efectos de los tratamientos, es decir, basándose en la uniformidad de las parejas. El emparejamiento de las unidades experimentales, por tanto, no es aleatorio, sino que es el criterio del experimentador el que decide cuáles son las parejas con arreglo a efectos ambientales o de clasificación; y dentro de cada pareja sí se aleatoriza el tratamiento que le corresponde a cada unidad experimental.

Es de esperar que los miembros de las parejas tiendan a correlacionarse positivamente, esto es, que los valores de las dos unidades experimentales de cada par tienden a ser grandes o pequeños conjuntamente, de manera que, por ejemplo, si un lechón pesa por encima de la media del peso de lechones, entonces su hermano también pesará por encima de la media. En este caso puede aumentarse la precisión del experimento para detectar pequeñas diferencias de la manera que se verá a continuación.

Si se va a probar, por ejemplo, dos dietas y para ello se realiza un experimento en el que se toman diez hermanos de sendas camadas de lechones dándole a cada hermano una de las dietas, al final del experimento se tendrá dos muestras, una con las diez medidas de la primera dieta ( $Y$ ) y la otra muestra con las diez medidas de la segunda dieta ( $X$ ). Se podría pensar que la prueba adecuada sería hacer un contraste de media de muestras independientes tal como la estudiada en el epígrafe anterior. En dicha prueba, el denominador del estadístico de prueba, la  $t_0$ , es el error típico de la diferencia de medias que implica la estima de la varianza conjunta,  $S^2_{1,2}$ , de ambas muestras mediante la suma de las dos varianzas. Por lo que es conveniente comprobar si la varianza de una suma es igual que la suma de las varianzas. No es muy complicado de comprobar, como se ha visto en el página 111.

Como se ha visto en dicha página, la expresión para los estadísticos muestrales de la varianza de la suma de dos variables es

$$S^2_{X_1+X_2} = S^2_{X_1} + S^2_{X_2} + 2 Cov$$

Esto indica que si se trabaja con una variable compuesta de la suma de otras dos variables, la varianza de esta nueva variable será la suma de las varianzas de las variables que la componen, más un nuevo término añadido que es función del grado en que estén asociados los valores de éstas dos variables. Este valor añadido a la suma de las dos varianzas es el doble de la covarianza.

Por lo que el error típico, para la prueba  $t$  de muestras independientes, es demasiado grande, al estar basado en la suma de las varianzas de las dos variables sin sustraerle ningún término debido a la covarianza.

Cuando las dos variables que se suman no están correlacionadas (no están asociadas, no varían conjuntamente), este término añadido de la covarianza será también cero y la varianza de la suma será simplemente la suma de las varianzas de las dos variables. Esta es la razón por la que se insiste en el supuesto de la *independencia* entre las dos variables a fin de poder sumar sus varianzas en el denominador del estadístico de prueba. De otro modo habría que haber tenido en consideración un término de covarianza que hasta ahora no se ha tenido en cuenta.

Sin embargo, si se tiene datos emparejados se espera encontrar correlación entre las variables, dado que cada pareja comparte una experiencia común.

De esta forma, siempre que se tenga dos medidas de una variable que estén asociadas positivamente, la varianza de sus diferencias será considerablemente menor que la suma de sus varianzas. Por esta razón, en casos como éste, se debe aplicar la prueba  $t$  para comparaciones de medias de muestras emparejadas en lugar de la prueba  $t$  para comparaciones de medias de muestras independientes.

El número de grados de libertad en que se basa la estimación de la varianza de las diferencias es el número de pares menos uno, mientras que los grados de libertad de la suma de las varianzas es de dos veces el número de pares menos dos, o lo que es lo mismo, la suma de los grados de libertad de ambas muestras (si fueran independientes) como se ha visto anteriormente. Por lo tanto, parece ser, se puede perder precisión en la prueba  $t$  de medias de muestras emparejadas al perder grados de libertad. Sin embargo no es así, sino que, por el contrario, se incrementa la capacidad de un experimento para detectar una diferencia real entre medias, porque la varianza de las diferencias ( $\sigma^2_D$ ) es mucho menor que la varianza conjunta ( $\sigma^2_{1,2}$ ) con lo que se compensa sobradamente la pérdida de grados de libertad debido al apareamiento.

Finalmente, qué ocurriría si existiera una correlación negativa. Si se tiene, siguiendo con el mismo ejemplo, parejas de animales tomados de sendas camadas pero, en este caso, se van a enjaular emparejados de manera que comparten la cantidad total de comida, que es constante para cada pareja (la aplicación de los tratamientos es manual e individualizada), podría ocurrir que si un animal come más, su pareja comerá menos por lo que al final de la experiencia se puede tener una correlación negativa en el aumento de peso. En tal caso, el error típico de la prueba de observaciones emparejadas será mayor que el de la prueba de muestras independientes, como consecuencia de que el término de la covarianza, al ser negativa, se suma a la adición de las dos varianzas. Por lo tanto, con una correlación negativa, una prueba  $t$  de observaciones emparejadas tiene menos probabilidad de rechazar la hipótesis nula que una prueba  $t$  de muestras independientes.

El criterio de prueba  $t$  es, como el que se presentó anteriormente,

$$t_o = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}} = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{D}}} = \frac{\bar{X}_D - \mu_D}{S_{\bar{D}}}$$

siendo

$$S_{\bar{D}} = \frac{S_D}{\sqrt{n}}$$

Es decir, el error típico de las diferencias, siendo la desviación típica de las diferencias

$$S_D = \sqrt{\frac{\sum_i (X_{1i} - X_{2i})^2 - \frac{[\sum (X_1 - X_2)]^2}{n}}{n-1}} = \sqrt{\frac{\sum_i D^2 - \frac{(\sum_i D_i)^2}{n}}{n-1}}$$

siendo  $n$  el número de diferencias de los pares de valores y, por lo tanto,  $gl = n-1$ .

Por lo que las pruebas de hipótesis con observaciones emparejadas se reducen a probar que la media de las diferencias es un número dado ( $\mu_D$ ). Lo más corriente es que se pruebe para

$$H_0 : \mu_D = 0 \quad H_0 : \mu_D \leq 0 \quad H_0 : \mu_D \geq 0$$

$$H_1 : \mu_D \neq 0 \quad H_1 : \mu_D > 0 \quad H_1 : \mu_D < 0$$

igualdad de efectos en la hipótesis nula.

Si se desea utilizar el intervalo de confianza, éste es

$$L_i = \bar{X}_D - t_{(n-1; \alpha/2)} S_{\bar{D}}$$

$$L^s = \bar{X}_D + t_{(n-1; \alpha/2)} S_{\bar{D}}$$

### Ejemplo.-

Se está tratando de aumentar la prolificidad media en cierta raza de ovejas neozelandesas. Para ello se han seguido dos tácticas en sendas líneas. Después de diez años de experimentación se desea saber si han divergido dichas líneas para el tamaño medio del número de descendencia. Los datos son

Año	Línea A	Línea B	D
1	1.70	1.38	+0.32
2	1.62	1.43	+0.19
3	1.45	1.41	+0.04
4	1.92	1.87	+0.05
5	1.96	1.84	+0.12
6	1.72	1.75	-0.03
7	1.68	1.60	+0.08
8	1.99	1.91	+0.08
9	1.87	1.80	+0.07
10	1.57	1.46	+0.11

$$\begin{aligned} \sum X_A &= 17.48 & \sum X_B &= 16.45 \\ \bar{X}_A &= 1.748 & \bar{X}_B &= 1.645 \\ S_A^2 &= 0.0325 & S_B^2 &= 0.0447 \\ n &= 10 & n &= 10 \end{aligned}$$

$$\begin{aligned} \sum X_A X_B &= 29.061 & \bar{X}_D &= 0.103 \\ SP &= 0.3064 & S_D^2 &= 0.0091 \\ Cov &= 0.03404 & S_{\bar{D}} &= 0.0301 \\ n &= 10 & n &= 10 \end{aligned}$$

Estos son datos emparejados, pues, además de los *tratamientos* (líneas), hay otro factor, los años, que hace que los datos se emparejen por años. Si pasa desapercibido este emparejamiento y se realiza una prueba *t* de muestras independientes, esta sería

Primeramente de prueba el supuesto paramétrico de la igualdad de varianzas

$$\begin{aligned} H_0 &: \sigma_1^2 = \sigma_2^2 \\ H_1 &: \sigma_1^2 \neq \sigma_2^2 \\ F_o &= \frac{0.0447}{0.0325} = 1.3754 ns \\ F_{(9,9; 0.05/2)} &= 4.03 \end{aligned}$$

Como la  $F_o$  es no significativa, se concluye que las varianzas de ambas muestras son estadísticamente iguales, por lo que se puede calcular la varianza conjunta

$$S_{1,2}^2 = \frac{9 \times 0.0447 + 9 \times 0.0325}{9 + 9} = 0.0386$$

se observa, en el denominador, que los grados de libertad de esta varianza es la suma de los grados de libertad de las dos muestras, es decir,  $g! = 18$ .

Y el error típico de la diferencia de medias sería

$$S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{2 \times 0.0386}{10}} = 0.0879$$

Por lo que ahora se puede hacer la prueba de igualdad de medias

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

$$t_o = \frac{17.48 - 16.45}{0.0879} = 1.1723ns$$

$$t_{(18, 0.05/2)} = 2.1009$$

Como  $t_o$  es no significativa, se concluye que las dos medias son iguales, es decir, que ambas líneas tienen la misma prolificidad.

Pero esta no es la prueba correcta pues, al ser datos emparejados, hay que eliminar la variabilidad *entre parejas*, es decir, la debida a los *años*, con la prueba  $t$  de las diferencias que se basa en el error típico de las diferencias.

Se calcula primero la desviación típica de las diferencias

$$S_D = \sqrt{\frac{0.1877 - \frac{1.03^2}{10}}{9}} = 0.0952$$

se observa, en el denominador, que los grados de libertad de esta desviación típica es el número de parejas menos uno, es decir,  $g! = 9$ .

El error típico de las diferencias, por tanto, es

$$S_{\bar{D}} = \frac{0.0952}{\sqrt{10}} = 0.0301$$

Con lo que se puede realizar la prueba  $t$  de las diferencias

$$H_0 : \mu_D = 0$$

$$H_1 : \mu_D \neq 0$$

$$t_o = \frac{0.1030}{0.0301} = 3.4205 **$$

$$t_{(9; 0.01/2)} = 3.2498$$

Como  $t_0$  es significativa, se concluye que las medias no son iguales, es decir, que ambas líneas tienen diferente tamaño medio de descendencia.

La diferencia en el resultado de ambas pruebas  $t$  indican la existencia de asociación entre los valores de ambas estirpes, o lo que es lo mismo, que el factor año influye significativamente en la variación de dichos valores, por lo que es conveniente tener esto en cuenta con objeto de que esta variabilidad entre parejas no encubra la variabilidad que interesa, esta es, la debida a las diferentes técnicas de selección que se han seguido con las dos líneas.

De acuerdo con las expresiones deducidas anteriormente, se puede calcular la  $Cov$  despejándola de dichas ecuaciones

$$Cov = \frac{S_1^2 + S_2^2 - S_D^2}{2} = \frac{0.0325 + 0.0447 - 0.0091}{2} = 0.03405$$

Como se ve, si estos datos se hubieran analizado mediante una prueba  $t$  para contraste de medias de muestras independientes, el resultado hubiera sido de no significación concluyéndose que las dos medias son iguales y que, por tanto, la prolificidad media sería igual en ambas líneas; cuando esto no es cierto. Lo que ha ocurrido es que esta prueba no ha tenido la precisión adecuada para detectar una diferencia real entre medias debido a que el error típico de esta prueba  $t$  es demasiado grande por estar basado en la suma de las varianzas de dos variables asociadas sin sustraerle ningún término debido a esta asociación, por lo que se está empleando para el contraste dos variabilidades, la debida al factor que interesa, el de las dos líneas o técnicas de selección, y la variabilidad entre años que es accesoria y que se sabía que tenía que existir, quedando solapada la variabilidad *entre las líneas* con la variabilidad de los *años*.

La prueba  $t$  para datos apareados sustrae automáticamente dicho término de variabilidad debido al otro factor (*años*), dando como resultado un error típico menor y, en consecuencia, un mayor valor de  $t$ , dado que el numerador es el mismo. Como consecuencia, se consigue una mayor precisión y se detecta las diferencias que la otra prueba no detectaba. Efectivamente, dicha prueba da significativa y se concluye que las medias son diferentes, o lo que es lo mismo, que el número medio de corderos nacidos en cada parto son diferentes para ambas líneas.

### Archivo del programa SAS (C4-6.SAS).-

```
title 'prueba t de observaciones emparejadas';
options ls=75 ps=60;
data p_t_empa;
infile 'c4-6.dat';
input ano lineaA lineaB;
dif = lineaA-lineaB;
proc means n mean var stderr t prt maxdec=4;
    var dif;
run;
```

### Archivo de datos (C4-6.DAT).-

1	1.70	1.38
2	1.62	1.43
3	1.45	1.41
4	1.92	1.87
5	1.96	1.84
6	1.72	1.75
7	1.68	1.60
8	1.99	1.91
9	1.87	1.80
10	1.57	1.46

### Archivo de resultados (C4-6.LST).-

N	Mean	Variance	Std Error \	T	Prob> T
10	0.1030	0.0091	0.0301	3.4205	0.0076

El archivo de resultados muestra que el valor de  $t$  es 3.4205, que es significativo al 0.01, por lo que concluye que han divergido ambas líneas en el tamaño medio del número de descendientes

### Modelo factorial sin repetición con un solo grado de libertad para los tratamientos.-

Se puede demostrar que el caso de comparaciones emparejadas puede ser considerado como un caso particular del análisis de varianza de dos factores sin repetición en el que uno de los factores es la variación entre pares. Este análisis es el de un diseño de bloques aleatorios.

Si se calcula la  $t$  de las diferencias entre los valores de cada par y se contrasta para una diferencia media de cero, se tiene que, en este caso, el cuadrado de la  $t_0$  será igual a la  $F_0$  del factorial sin repetición.

### Archivo del programa SAS (C4-7.SAS).-

```
title 'observaciones emparejadas. Análisis de la varianza, dos factores sin
repetición con 1 grado de libertad';
options ls=75 ps=60;
data pru_F;
infile 'c4-7.dat';
input ano linea $ hijos @@;
proc anova;
  class ano linea;
  model hijos = ano linea;
run;
```

### Archivo de datos (C4-7.DAT).-

1	A	1.70	1	B	1.38
---	---	------	---	---	------

2 A 1.62	2 B 1.43
3 A 1.45	3 B 1.41
4 A 1.92	4 B 1.87
5 A 1.96	5 B 1.84
6 A 1.72	6 B 1.75
7 A 1.68	7 B 1.60
8 A 1.99	8 B 1.91
9 A 1.87	9 B 1.80
10 A 1.57	10 B 1.46

### Archivo de resultados (C4-7.LST).-

Dependent Variable: HIJOS					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	0.70665000	0.07066500	15.59	0.0002
Error	9	0.04080500	0.00453389		
Corrected Total	19	0.74745500			
R-Square		C.V.	Root MSE	HIJOS Mean	
0.945408		3.969004	0.0673342	1.6965000	
Source	DF	Anova SS	Mean Square	F Value	Pr > F
ANO	9	0.65360500	0.07262278	16.02	0.0002
LINEA	1	0.05304500	0.05304500	11.70	0.0076

El archivo de resultados muestra que el valor de  $F$  para LINEA es exactamente el cuadrado de la  $t$  anterior y que tiene exactamente la misma significación, por lo que son equivalentes ambos análisis.

### Confusión de modelos.-

Hay que hacer notar que si estos datos se hubieran analizado como si fueran dos muestras independientes hubiera dado no significativa, esto es, que el tamaño medio de descendientes es el mismo

### Archivo del programa SAS (C4-8.SAS).-

```

title 'Prueba t de observaciones emparejadas como si fueran muestras
independientes';
options ls=75 ps=60;
data p_t_empa;
infile 'c4-8.dat';
input ano linea $ hijos @@;
proc ttest;
  class linea;
  var hijos;
run;

title 'Análisis de varianza de observaciones emparejadas como una vía, 2
tratamientos';
options ls=75 ps=60;
proc anova;
  class linea;
  model hijos = linea;
run;

```



**Archivo de datos (C4-8.DAT).-**

A	1.70	B	1.38
A	1.62	B	1.43
A	1.45	B	1.41
A	1.92	B	1.87
A	1.96	B	1.84
A	1.72	B	1.75
A	1.68	B	1.60
A	1.99	B	1.91
A	1.87	B	1.80
A	1.57	B	1.46

**Archivo de resultados (C4-8.LST).-**

TTEST PROCEDURE						
LINEA	N	Mean	Std Dev	Std Error	Minimum	Maximum
A	10	1.74800000	0.18029605	0.05701462	1.45000000	1.99000000
B	10	1.64500000	0.21130547	0.06682066	1.38000000	1.91000000
Variances		T	DF	Prob> T		
Unequal		1.1726	17.6	0.2566		
Equal		1.1726	18.0	0.2562		
For H0: Variances are equal, F. = 1.37    DF = (9,9)    Prob>F. = 0.6440						
Analysis of Variance Procedure						
Dependent Variable: HIJOS						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	1	0.05304500	0.05304500	1.37	0.2562	
Error	18	0.69441000	0.03857833			
Corrected Total	19	0.74745500				
R-Square	C.V.	Root MSE	HIJOS Mean			
0.070967	11.57758	0.1964137	1.6965000			
Source	DF	Anova SS	Mean Square	F Value	Pr > F	
LINEA	1	0.05304500	0.05304500	1.37	0.2562	

Como se ha visto anteriormente, la  $t$  es la raíz cuadrada de la  $F$ , y ambas dan el mismo nivel de significación porque son redundantes, estos son  $t = 1.173ns$  para  $gl = 18$  y  $F = 1.07ns$  para  $gl_{numerador} = 1$ ,  $gl_{denominador} = 18$ ; por lo que ambas medias son iguales

**Generalización multivariante.-**

En el caso de más de una variable, la  $T^2$  de *Hotteling* se puede calcular igual que se hizo cuando se tenía una sola muestra, ésta era

$$T^2 = F \frac{V(N-1)}{N-V}$$

siendo  $V$  el número de variables, y siendo  $N$  el número de pares de datos.

### Ejemplo.-

Se quiere saber los efectos de cierto fármaco sobre el nivel de glucosa en sangre así como sobre la presión sanguínea, para ello se mide estas tres variables a ocho personas, después se les inocula 2.0 mg/Kg de dicho fármaco y se les vuelve a medir las mismas variables, dando los siguiente resultados

	Antes			Después		
	Gluco	P. sistó	P. diastó	Gluco	P. sistó	P. diastó
1	90	148	79	120	140	78
2	60	123	80	150	130	86
3	105	132	84	95	130	88
4	100	121	85	110	121	87
5	65	134	83	95	132	88
6	90	121	82	150	121	85
7	85	123	89	85	121	93
8	72	110	87	112	111	89

### Archivo del programa SAS (C4-9.SAS).-

```

title 'observaciones emparejadas. Análisis de la varianza, dos factores sin
repetición con 1 grado de libertad. Multivariante';
data prut2;
options ls=75 ps=60;
infile 'c4-9.dat';
input pobla $ indi glucosa psisto pdiasto;
proc glm;
class pobla indi;
model glucosa psisto pdiasto = pobla indi;
manova h=pobla indi;
run;

```

### Archivo de datos (C4-9.DAT).-

```

Antes 1 90 148 79
Antes 2 60 123 80
Antes 3 105 132 84
Antes 4 100 121 85
Antes 5 65 134 83
Antes 6 90 121 82
Antes 7 85 123 89
Antes 8 72 110 87
Despues 1 120 140 78
Despues 2 150 130 86
Despues 3 95 130 88
Despues 4 110 121 87
Despues 5 95 132 88
Despues 6 150 121 85
Despues 7 85 121 93

```

**Archivo de resultados (C4-9.LST).-**

Las primeras salidas son las de los anova univariantes. Se puede comprobar por los valores de las  $F$  significativos que ha habido cambios en la glucosa y en la presión diastólica, pero no en la presión sistólica.

La salida del análisis multivariante para el factor **población** es significativa, lo que indica que existe una diferencia significativa entre la respuesta media de las tres variables antes y después de la administración del fármaco. Las diferencias entre los individuos también son significativas para las tres variables consideradas en conjunto.

Manova Test Criteria and Exact F Statistics for the Hypothesis of no Overall POBLA Effect					
H = Type III SS&CP Matrix for POBLA    E = Error SS&CP Matrix					
	S=1	M=0.5	N=1.5		
Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks. Lambda	0.08133519	18.8247	3	5	0.0037
Pillai's Trace	0.91866481	18.8247	3	5	0.0037
Hotelling-Lawley Trace	11.29480108	18.8247	3	5	0.0037
Roy's Greatest Root	11.29480108	18.8247	3	5	0.0037

El valor de la  $T^2$  de Hotelling es

$$T^2 = 18.8247 \frac{3(8-1)}{8-3} = 79.0637$$

**Caso especial: un solo ejemplar comparado con una muestra.-**

Este caso se puede considerar como uno especial del primer caso en el que una de las dos muestras contiene una única observación, de manera que no se puede calcular la varianza de esta.

El estadístico de prueba adecuada en este caso sería

$$t_o = \frac{X_1 - \bar{X}_2}{S_2 \sqrt{\frac{n_1-1}{n_2}}}$$

Que se contrastaría con la  $t$  de para  $gl = n_2 - 1$

## Comparación de dos medias de muestras independientes y varianzas desiguales.-

Algunas veces puede ocurrir que al aplicar, o al venir dados, los efectos de dos tratamientos sobre una misma variable, se generan dos poblaciones o distribuciones normales diferentes (diferentes varianzas y diferentes medias). Si se quiere conocer la diferencia real entre los efectos de estos dos tratamientos, se tendría que conocer las dos poblaciones y comprobar si están situadas en el mismo punto, es decir, tienen la misma media (no existe diferencia en los efectos) o, por el contrario, una está desplazada a la derecha de la otra en el eje de las abscisas (el efecto de un tratamiento es mayor que el efecto del otro tratamiento). Como no es posible conocer las poblaciones al ser infinitas, se tendrá que hacer, como siempre, tomando muestras y a partir de ellas inferir (*estimar*) sobre el parámetro de posición de ambas poblaciones, es decir, sobre la diferencia de las medias.

Para comparar  $\mu_1$  y  $\mu_2$ , se toma una muestra de tamaño  $n_1$  de la primera población y una muestra de tamaño  $n_2$  de la segunda población, teniendo, por tanto

$$\begin{aligned} X_{11}, X_{12}, \dots, X_{1n_1} &\sim N(\mu_1, \sigma_1^2) \\ X_{21}, X_{22}, \dots, X_{2n_2} &\sim N(\mu_2, \sigma_2^2) \end{aligned}$$

Se desea saber si las medias de las dos poblaciones son iguales, es decir, si  $\mu_1 = \mu_2$ .

Sean los dos parámetros de cada población,  $\mu_1$ ,  $\mu_2$  y  $\sigma_1^2$ ,  $\sigma_2^2$  las medias y las varianzas de dichas poblaciones. Para comparar  $\mu_1$  y  $\mu_2$ , hay que hacerlo a través de sus medias muestrales o estimas respectivas,  $\bar{X}_1$  y  $\bar{X}_2$ .

Se puede probar si las medias de ambas poblaciones son iguales, contrastando

$$\begin{aligned} H_0 : \mu_1 - \mu_2 &= 0 \\ H_1 : \mu_1 - \mu_2 &\neq 0 \end{aligned}$$

para ello se utiliza las diferencias de las estimas  $\bar{X}_1 - \bar{X}_2$ .

Esta claro que un valor de  $\bar{X}_1 - \bar{X}_2$  cercano a cero apoyara la hipótesis nula (medias iguales), mientras que un valor alejado del cero apoyará la hipótesis alternativa (medias diferentes).

Ya se ha demostrado anteriormente, en este mismo que la media de una diferencia es igual a la diferencia de las medias

$$\mu_{(X_1 - X_2)} = \mu_1 - \mu_2$$

y que la varianza de una diferencia es igual

$$\sigma_{(X_1 - X_2)}^2 = \sigma_{X_1}^2 + \sigma_{X_2}^2 - 2\sigma_{X_1 X_2}$$

siendo  $\sigma_{X_1 X_2}$  la covarianza paramétrica. Si las muestras son independientes, y por lo tanto no existe asociación entre los valores de una y otra población, este término vale cero, por lo que la varianza de una diferencia queda

$$\sigma_{(X_1 - X_2)}^2 = \sigma_{X_1}^2 + \sigma_{X_2}^2$$

Por tanto, la distribución muestral de  $\bar{X}_1 - \bar{X}_2$  es

$$\bar{X}_1 - \bar{X}_2 \sim N \left( \mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right)$$

siendo

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

el error típico de la diferencia de medias.

Como los valores de  $\bar{X}_1 - \bar{X}_2$  tenderán a estar concentrados, con probabilidad alta, alrededor del valor  $\mu_1 - \mu_2$ , se tiene que la distribución muestral del estadístico

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{X}_1 - \bar{X}_2}}$$

es una normal típica, es decir, media cero y varianza uno. Por lo que una regla de decisión razonable para probar las hipótesis planteadas sería

$$\text{Rechazar } H_0 \text{ si } Z_0 = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}} \geq Z_{(\alpha)}$$

siendo  $Z_\alpha$  el valor de  $Z$  para el nivel de significación elegido, habitualmente para  $\alpha=0.05$ ,  $\alpha=0.01$  y  $\alpha=0.001$ .

Pero para hacer uso de esta prueba hay que conocer los valores de los parámetros  $\sigma_1^2$  y  $\sigma_2^2$ , situación poco frecuente. Lo habitual es que se desconozca el valor de las varianzas y que, por tanto, haya que usar sus estimas a través de los estadísticos muestrales, es decir  $S_1^2$  y  $S_2^2$ .

Por tanto, se usará como error típico de la diferencia de medias la estima siguiente

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Como se tienen dos distribuciones diferentes (varianzas diferentes), ni la suma de cuadrados ni los grados de libertad se combinan como cuando decir  $S_1^2$  y  $S_2^2$  eran estimas de una varianza común  $\sigma^2$ . Por tanto, los  $gl$  será un valor entre  $n_1$  y  $n_2$ , inclusive.

Más adelante de verán dos aproximaciones para el cálculo de los grados de libertad de este error típico.

La distribución muestral del estadístico queda como sigue

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_{\bar{X}_1 - \bar{X}_2}}$$

De nuevo,  $t$  mide la distancia de una variable aleatoria o una media hipotética en unidades de desviación típica de la variable aleatoria, por lo que su distribución es como la  $t$  de *Student*.

En este caso, una regla de decisión razonable para probar las hipótesis planteadas sería

$$\text{Rechazar } H_0 \text{ si } t_o = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}} \geq t_{(n; \alpha/2)}$$

siendo  $t_{(n; \alpha/2)}$  el valor de  $t$  (tabla 2) para  $gl$  igual a un valor intermedio entre  $n_1$  y  $n_2$ , y el nivel de significación elegido, habitualmente para  $\alpha=0.05$ ,  $\alpha=0.01$  y  $\alpha=0.001$ .

Se han desarrollado varias aproximaciones para resolver la prueba  $t$  cuando las varianzas son diferentes. Aquí se van a presentar la de *Satterthwaite* (*Steel y Torrie, 1985*) y la de *Cochran y Cox* (*Snedecor y Cochran, 1971*).

### **Aproximación de Cochran y Cox.-**

En la aproximación de *Cochran y Cox* se vuelve a hacer una subdivisión en dos casos.

### **Comparación de medias cuando $\sigma_1^2 \neq \sigma_2^2$ y $n_1 = n_2 = n$ .-**

En este caso, no existe duda con respecto a los  $gl$ , puesto que al ser las dos muestras del mismo tamaño, el error típico de la diferencia de medias es

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{n}} = \sqrt{\frac{S_1^2 + S_2^2}{n}}$$

y, por lo tanto  $gl = n-1$ .

Si se multiplica y divide por 2 el numerador de la raíz cuadrada del valor del error típico de la diferencias de medias, este no cambia

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{2 \frac{S_1^2 + S_2^2}{2}}{n}} = \sqrt{\frac{2 S_{1,2}^2}{n}}$$

puesto que

$$S_{1,2}^2 = \frac{(n-1)S_1^2 + (n-1)S_2^2}{(n-1) + (n-1)} = \frac{(n-1)(S_1^2 + S_2^2)}{2(n-1)} = \frac{S_1^2 + S_2^2}{2}$$

Que no es sino el error típico de la diferencia de medias que se calculó cuando las muestras son independientes, del mismo tamaño y las varianzas iguales; por lo que el cálculo de  $t$  es el mismo que en ese caso, lo que cambia es el valor crítico con el que va a contrastar. En este caso se contrasta, como se ha deducido más arriba, con  $t_{(n-1; \alpha/2)}$ , en lugar de  $t_{(2(n-1); \alpha/2)}$ .

### Comparación de medias cuando $\sigma_1^2 \neq \sigma_2^2$ y $n_1 \neq n_2$ -

En este caso se calcula la  $t$  con el error típico de la diferencia de medias que se ha deducido más arriba, es decir

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Como hay dos tamaños de muestra, el problema es saber con cuál  $t$  de las tablas se contrasta la  $t_o$ , si con  $t_{(n_1-1; \alpha/2)} = t_1$  o con  $t_{(n_2-1; \alpha/2)} = t_2$ , o lo que es lo mismo, ¿con cuántos grados de libertad se mira en las tablas?

En este caso hay que estimar una  $t$  teórica ponderada para la magnitud de las varianzas y del tamaño de muestra. Esta  $t_p$  es

$$t_p = \frac{\left( t_1 \frac{S_1^2}{n_1} \right) + \left( t_2 \frac{S_2^2}{n_2} \right)}{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

siendo

$$t_1 = t_{(n_1-1; \alpha/2)} \quad \text{y} \quad t_2 = t_{(n_2-1; \alpha/2)}$$

## Aproximación de *Satterthwaite*.-

En la aproximación de *Satterthwaite* no hay subdivisión en dos casos, según los tamaños de muestra, sino que a ambos casos se le aplica la misma solución, que viene a ser la misma del caso 2 de la aproximación de *Cochran* y *Cox* que se acaba de exponer, con la única variación de lo que se calcula son los *gl ponderados*, también para la magnitud de las varianzas y para los tamaños de muestra (en lugar de la *t ponderada*) de la siguiente manera

$$g^l_p = \frac{\left( \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\frac{\left( \frac{S_1^2}{n_1} \right)^2}{n_1-1} + \frac{\left( \frac{S_2^2}{n_2} \right)^2}{n_2-1}}$$

Es de esperar que estos  $g^l_p$  no sea un número entero por lo que hay que redondearlo.

El criterio por intervalos de confianza es

$$L_i = \bar{X}_1 - \bar{X}_2 - \left[ t_{p(\alpha/2)} S_{\bar{X}_1 - \bar{X}_2} \right]$$
$$L^s = \bar{X}_1 - \bar{X}_2 + \left[ t_{p(\alpha/2)} S_{\bar{X}_1 - \bar{X}_2} \right]$$

O bien

$$L_i = \bar{X}_1 - \bar{X}_2 - \left[ t_{(g^l_p; \alpha/2)} S_{\bar{X}_1 - \bar{X}_2} \right]$$
$$L^s = \bar{X}_1 - \bar{X}_2 + \left[ t_{(g^l_p; \alpha/2)} S_{\bar{X}_1 - \bar{X}_2} \right]$$

El paquete estadístico *SAS* ha adoptado ambas aproximaciones, en la opción **COCHRAN** del **PROC TTEST**. De todas formas hay que hacer notar que las dos soluciones dan resultados finales semejantes.

## Ejemplo.-

Se está estudiando el porcentaje de la fracción lipoproteica del suero en vacuno de leche al inicio y al final de la lactación, con objeto de conocer si éste porcentaje es el mismo en ambos periodos del ciclo reproductivo. Los datos son



<i>Inicio</i>	<i>Final</i>
51.30	12.85
58.32	18.30
41.93	68.00
56.93	42.00
46.42	58.20
48.99	23.23
45.41	17.43
58.00	13.20
57.90	56.92
46.50	27.02

$$n_1 = 10 \qquad n_2 = 10$$

$$\bar{X}_1 = 51.17 \qquad \bar{X}_2 = 33.715$$

$$S_1^2 = 38.2153 \qquad S_2^2 = 432.8761$$

Al ser la variable porcentajes, puede ocurrir fácilmente que se tengan dos distribuciones diferentes, esto es, que tengan diferentes la varianza, por lo que se comprueba, primeramente, si las varianzas son iguales o no

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

$$F_o = \frac{432.8761}{38.2153} = 11.3273^{**}$$

$$F_{(9, 9; 0.01/2)} = 6.54$$

se rechaza la hipótesis nula, por lo que, efectivamente, son distintas las dos varianzas.

Se hace, entonces, la prueba de hipótesis de la diferencias de medias en el porcentaje de la fracción de lipoproteína en los dos periodos

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 \neq 0$$

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{38.2153}{10} + \frac{432.8761}{10}} = 6.8636$$

$$t_o = \frac{51.17 - 33.715}{6.8636} = 2.5431^*$$

*Aproximación de Cochran y Cox*

$$t_{(9; 0.05/2)} = 2.2622$$

### Aproximación de Satterthwaite

$$gI_p = \frac{\left(\frac{38.2152}{10} + \frac{432.8761}{10}\right)^2}{\frac{\left(\frac{38.2152}{10}\right)^2}{9} + \frac{\left(\frac{432.8761}{10}\right)^2}{9}} = 10.6 \approx 11$$

$$t_{(11;0.05/2)} = 2.2010$$

La conclusión es la misma con las dos aproximaciones. Se rechaza la hipótesis nula, por lo que se concluye que la concentración de lipoproteínas es diferente en los dos periodos. Efectivamente, hay una mayor concentración de lipoproteína al comienzo de la lactación que al final.

### Archivo del programa SAS (C4-10.SAS)-

```
title 'prueba t, muestras independientes, varianzas desiguales';
options ls=80 ps=60;
data pruebat;
infile 'c4-10.dat';
input periodo $ lipo @@;
proc ttest cochran;
class periodo;
var lipo;
run;
```

### Archivo de datos (C4-10.DAT)-

```
I 51.30 F 12.85
I 58.32 F 18.30
I 41.93 F 68.00
I 56.93 F 42.00
I 46.42 F 58.20
I 48.99 F 23.23
I 45.41 F 17.43
I 58.00 F 13.20
I 57.90 F 56.92
I 46.50 F 27.02
```

En el archivo de resultados (**C4-10.LST**) se observa que las varianzas son desiguales, pues la  $F = 11.33$  es significativa al  $P < 0.0013$ .

El valor de  $t$  para varianzas desiguales es de  $-2.5431$  que es significativo tanto con la aproximación de Satterthwaite (con  $gI = 10.6$ ) como con la aproximación de Cochran (con  $gI = 9$ ).

PERIODO	N	Mean	Std Dev	Std Error	Minimum	Maximum
F	10	33.71500000	20.80567457	6.57933199	12.85000000	68.00000000
I	10	51.17000000	6.18185337	1.95487368	41.93000000	58.32000000
Variances		T	Method	DF	Prob> T	
Unequal	-2.5431	Satterthwaite	10.6	0.0281		
		Cochran	9.0	0.0315		
Equal	-2.5431		18.0	0.0204		
For H0: Variances are equal, F. = 11.33 DF = (9,9) Prob>F. = 0.0013						

### Importancia de los supuestos paramétricos.-

Pudiera ocurrir que, si no se tienen en cuenta los supuestos paramétricos, las conclusiones a las que lleguemos no sean del todo correctas

### Ejemplo.-

Se está probando un nuevo método analítico más rápido para detectar la concentración de cierto compuesto químico. Para ello se hacen ocho determinaciones con este nuevo método, mientras que hacemos cuatro con el viejo método, que servirán de control o testigo. Los datos son

Método viejo: 25, 24, 25, 26  
Método nuevo: 23, 18, 22, 28, 17, 25, 19, 16

$$n_1 = 4 \qquad n_2 = 4$$

$$\bar{X}_1 = 25 \qquad \bar{X}_2 = 21$$

$$S_1^2 = 0.6666 \qquad S_2^2 = 17.7143$$

Al ser dos métodos diferentes, puede ocurrir fácilmente que se tengan dos distribuciones diferentes, tanto para la media como para la varianza, por lo que se comprueba, primeramente, si las varianzas son iguales o no

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

$$F_o = \frac{17.7143}{0.666} = 26.5714 *$$

$$F_{(7, 3; 0.05/2)} = 14.6$$

se rechaza la hipótesis nula, por lo que, efectivamente, son distintas las dos varianzas.

Se hace, entonces, la prueba de hipótesis de la diferencias de medias en la cantidad de este compuesto químico

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{0.6666}{4} + \frac{17.7143}{8}} = 1.5430$$

$$t_o = \frac{25 - 21}{1.5430} = 2.5923^*$$

*Aproximación de Cochran y Cox*

$$t_p = \frac{\left(3.1825 \frac{0.6666}{4}\right) + \left(2.3646 \frac{17.7143}{8}\right)}{\frac{0.6666}{4} + \frac{17.7143}{8}} = 2.4218$$

*Aproximación de Satterhwaite*

$$gI_p = \frac{\left(\frac{0.6666}{4} + \frac{17.7143}{8}\right)^2}{\frac{\left(\frac{0.6666}{4}\right)^2}{3} + \frac{\left(\frac{17.7143}{8}\right)^2}{7}} = 7.99 \approx 8$$

$$t_{(8; 0.05/2)} = 2.3060$$

La conclusión es la misma con las dos aproximaciones. Se rechaza la hipótesis nula, por lo que se concluye que la concentración del producto químico detectado por los dos métodos es diferentes. El nuevo método subestima las concentraciones del elemento químico.

Obsérvese que si el análisis se hubiera hecho como si las dos varianzas fueran iguales, la conclusión sería que los dos métodos son iguales, por lo que se podría adoptar el nuevo método cuando realmente este método mide por debajo del valor correcto.

### Archivo del programa SAS (C4-11.SAS)-

```

title 'Prueba t de muestras independientes varianzas desiguales tamaño
diferente';
option ls=75 ps=60;
data pruebat;
infile 'c4-11.dat';
input metodo $ concentr @@;
proc ttest cochran;
  class metodo;
  var concentr;
run;

```

### Archivo de datos (C4-11.DAT).-

V 25	N 23
V 24	N 18
V 25	N 22
V 26	N 28
	N 17
	N 25
	N 19
	N 16

### Archivo de resultados (C4-11.LST).-

METODO	N	Mean	Std Dev	Std Error	Minimum	Maximum
N	8	21.00000000	4.20883425	1.48804762	16.00000000	28.00000000
V	4	25.00000000	0.81649658	0.40824829	24.00000000	26.00000000
Variances		T	Method	DF	Prob> T	
Unequal	-2.5923		Satterthwaite	8.0	0.0320	
			Cochran	.	0.0396	
Equal	-1.8402			10.0	0.0956	
For H0: Variances are equal, F. = 26.57    DF = (7,3)    Prob>F. = 0.0212						

### Elección del tamaño de muestra en comparaciones de dos o más muestras.-

En el Capítulo anterior se vio la elección del tamaño de muestra cuando lo que se quiere es estimar una media con una diferencia  $\delta$  del parámetro  $\mu$ , ahora se va a ver el tamaño de muestra requerido para que

$$\delta \geq |\mu_1 - \mu_2|$$

Al diseñar un experimento, además de establecer claramente los objetivos como preguntas que han de ser respondidas, esto es, las hipótesis que han de probarse y los efectos que han de estimarse, hay que determinar la precisión o cantidad de información que se va a obtener.

La precisión o cantidad de información, se mide con la inversa de la varianza de la media. Si  $I$  representa la cantidad de información, entonces

$$I = \frac{1}{\sigma_{\bar{X}}^2} = \frac{n}{\sigma^2}$$

A medida que  $\sigma^2$  aumenta, la cantidad de información disminuye, o a medida que aumenta  $n$  (tamaño de muestra) la cantidad de información aumenta.

Si se trata de un diseño experimental, cuando un tratamiento aparece  $n$  veces en un experimento se dice que el tratamiento está repetido.

Cualquiera que sea la fuente de los errores experimentales, el aumento del tamaño de muestra disminuye constantemente el error asociado a la diferencia entre los resultados medios de dos tratamientos, siempre y cuando se hayan tomado algunas precauciones (tales como la aleatorización) para asegurar que un tratamiento no está más favorecido en una repetición que en otra, de tal forma que los errores que afecten a cualquier tratamiento tiendan a anularse cuando los tamaño de muestras aumentan.

Los objetivos para que se midan  $n$  individuos en cada tratamiento son:

- 1) Permitir una estimación del error experimental.
- 2) Mejorar la precisión de un experimento mediante la reducción de la desviación típica de la media de un tratamiento.
- 3) Aumentar el alcance de la inferencia del experimento a través de la selección y del uso apropiado de unidades experimentales más variables.
- 4) Controlar la varianza del error.

Tanto para las pruebas de significación como para la estima de intervalos de confianza, es necesario estimar el error experimental.

En un experimento en el cual cada tratamiento aparece sólo una vez, no es posible estimar el error experimental. Si se observa una diferencia en este caso puede explicarse como una diferencia entre tratamientos o una diferencia debida a la variabilidad individual entre unidades experimentales por lo que es imposible saber con seguridad cuál explicación es la correcta. Es decir, cuando no existe un método para estimar el error experimental, no hay manera de determinar si las diferencias observadas indican diferencias reales entre los tratamientos o si se deben a la variabilidad individual teniendo como consecuencia que el experimento no es autosuficiente.

A medida que el número de repeticiones aumenta, las estimas de las medias poblacionales se hacen más precisas, esto es. Si se detecta una diferencia de cinco unidades usando cuatro repeticiones, un experimento de aproximadamente 16 repeticiones detectará la mitad de esa diferencia, o sea, 2.5 unidades.

La magnitud básica que se usa para medir errores experimentales es la *varianza del error*, es decir, la *esperanza matemática del cuadrado medio del error*, que afecta a la observación de una simple unidad experimental. La raíz cuadrada de esta cantidad se llama *error típico por unidad*. Si tenemos  $\sigma^2$  que es la varianza del error por unidad experimental, y hay  $n$  repeticiones, la varianza del error de la diferencia entre las medias de dos tratamientos es

$$\frac{2\sigma^2}{n}$$

y el error típico correspondiente es

$$\sqrt{\frac{2\sigma^2}{n}}$$

En ciertos tipos de experimentos, la repetición es un medio de aumentar el alcance de la inferencia de un experimento. La población muestreada es menos restringida en su definición y así se amplía la inferencia. Por ejemplo, supóngase que deseamos determinar si existe una diferencia real en el rendimiento lechero de dos razas de cierta especie en una región, y que hay dos tipos principales de explotación (intensiva, extensiva) en esta región. Si el objetivo del experimentador es obtener inferencias para ambos tipos de explotaciones, es obvio que ambas deben entrar en el experimento. También es importante que el área incluida dentro de cada repetición, esto es, en cada par de unidades experimentales (sendas ganaderías) en las cuales se tienen las dos razas, sea de un tipo de explotación tan uniforme como sea posible. No es necesario, y puede ser indeseable, tener condiciones muy uniformes entre repeticiones, especialmente si la población en cuestión es bastante amplia.

En muchas experiencias sobre el terreno, el experimento se repite en un periodo de varios años. La razón es obvia, ya que las condiciones varían año a año y es importante conocer el efecto de los años sobre las diferencias entre los tratamientos, porque las recomendaciones se suelen hacer para años futuros. De la misma manera, se usan diferentes localidades para evaluar los tratamientos en las diferentes condiciones ambientales presentes en la población, es decir, en la región para la cual van a hacerse las recomendaciones. Tanto las repeticiones en tiempo (*año*) como en espacio (*localidades*) pueden considerarse tipos amplios de repeticiones. El propósito es aumentar el alcance de la inferencia. Este mismo principio se usa en experimentos de laboratorio, o sea, que el experimento completo se repita varias veces, posiblemente por personas diferentes, para determinar la repetibilidad de los tratamientos en las diferentes condiciones que puedan existir en el laboratorio de tiempo en tiempo. Esta cuestión la trataremos más detenidamente en el siguiente capítulo.

Finalmente, la repetición nos permite agrupar unidades experimentales de acuerdo con la respuesta esperada en ausencia de tratamientos. El objeto es repartir la variación total entre unidades experimentales de tal manera que se maximice entre los grupos y se minimice simultáneamente dentro de ellos. Ahora bien, si se aplica un conjunto de tratamientos dentro de cada grupo de éstos, las diferencias observadas medirán las verdaderas diferencias de tratamientos mejor que si la agrupación se hubiera pasado por alto. Es claro que debe procurarse no incrementar el error experimental con objeto de detectar diferencias reales entre tratamientos, debido a diferencias entre grupos. En el siguiente capítulo se tratará más detenidamente.

Por lo tanto, el número de repeticiones de un experimento o tamaño de muestra de un muestreo depende de varios factores, de los cuales el más importante es el grado de precisión deseada. Cuanto más pequeña sea la discrepancia con respecto a la hipótesis nula que se desea medir o detectar, mayor será el número de repeticiones requeridas.

En un experimento o muestreo, es muy importante tener un grado correcto de precisión. No tiene objeto muestrear diez individuos para detectar una diferencia que se puede detectar con cuatro individuos. Tampoco es de utilidad llevar a cabo un experimento en el que el tamaño de muestra sea insuficiente para detectar diferencias importantes.

Para medir cualquier discrepancia con respecto a la hipótesis nula, el error experimental, esto es, la variación entre observaciones experimentales tratadas de manera análoga, debe proporcionarnos la unidad de medida para el error. A veces no es práctico hacer una observación en la unidad experimental completa, por ejemplo, en el caso de determinaciones químicas, como la del contenido proteínico de un forraje, se toman muestras de la unidad experimental. Usualmente la variación entre unidades experimentales es grande en comparación con la variación entre muestras de una sola unidad por lo que no hay motivo para realizar un número grande de determinaciones químicas en cada unidad experimental, ya que el error experimental debe estar basado en la variación entre las unidades experimentales, no en la variación entre las muestras (*unidad muestra*) de esas unidades experimentales.

Ciertos materiales experimentales son más variables que otros. Por ejemplo, ciertas razas de cierta ganadería son más uniformes que otras razas, por lo que, para conseguir una misma precisión, en las razas que son más homogéneas se necesitan tamaños de muestra menores que en las razas más heterogéneas.

El número de poblaciones (número de tratamientos) afecta la precisión de un experimento y los tamaños de muestras necesarios para un grado de precisión dado. Por ejemplo, si aumentamos el número de poblaciones y mantenemos constante el tamaño de muestra para cada uno, entonces aumentamos el tamaño del experimento y el número de grados de libertad para la estima de  $\sigma^2$ . La varianza de la media ( $S^2/n$ ) sigue siendo la misma pero mejora la estima de  $\sigma^2$  y, por tanto, es mayor la precisión. Por lo que en este caso, si no se requiere un aumento de precisión, se puede reducir el tamaño de muestra de cada población.

Por otra parte, si mantenemos constante el tamaño del experimento, entonces un mayor número de tratamientos implica menor tamaño de muestra de cada población y menos grados de libertad para estimar  $\sigma^2$ . Ahora  $S^2/n$  tiene un menor  $n$  y una estima más deficiente de  $\sigma^2$ . Como resultado se tiene menor precisión y tendría que aumentarse los tamaños de muestra para lograr una precisión fijada. Todo este razonamiento es más apropiado para experimentos pequeños, por ejemplo, con menos de 20 grados de libertad para el error.

El diseño experimental también afecta la precisión de un experimento y el número de repeticiones necesarias. Cuando el número de tratamientos es grande y es necesario usar unidades experimentales más heterogéneas, entonces aumenta el error experimental por unidad. Los diseños experimentales apropiados pueden controlar parte de esta variación.

Por tanto, la probabilidad de obtener un resultado significativo depende del verdadero error típico de la unidad experimental,  $\sigma$ , del tamaño de muestra  $n$  y de los grados de libertad del error.

El cálculo exacto de esta probabilidad es algo complicado pero existen aproximaciones bastante buena para propósitos prácticos. Una de ellas puede ser la del siguiente ejemplo.



### Ejemplo.-

Considérese la diferencia entre los efectos de un par de tratamientos, que pueden ser un procedimiento estándar y un nuevo procedimiento que se espera sea mejor. El error típico verdadero por unidad es del 12% de la media, hay ocho repeticiones y 14 grados de libertad para la estima de la varianza del error ¿Cuál es la probabilidad de obtener un resultado significativo cuando la diferencia verdadera es del 20%?.

Sea  $d$  la diferencia observada y  $\delta$  la diferencia verdadera, entre las medias del nuevo tratamiento y del tratamiento estándar. El error típico verdadero de  $d$  es

$$\sqrt{\frac{2\sigma^2}{n}} = \sqrt{\frac{2(12)^2}{8}} = 6\%$$

En la práctica no se conoce  $\sigma$ , y en su lugar se usa el error típico estimado por unidad  $S$ , obtenido del análisis de la varianza. Se supone que  $S$  es igual a  $\sigma$ .

La prueba de significación de  $d$  se hace por medio de una  $t$  de *Student*, con  $g=14$ . Para encontrar un valor de  $d$  que sea significativo, se multiplica el error típico de  $d$ , 6%, por el valor significativo de  $t$ . Si la prueba se va a hacer con  $\alpha=0.05$  y se trata de una prueba de dos colas, el valor al 10% de  $t_{(14;0.1)}$  es de 1.7613, por tanto, si  $d$  va a ser significativa, se debe tener

$$d \geq \sqrt{\frac{2S^2}{n}} t_1 = 6 \times 1.7613 = 10.5678$$

El problema está en encontrar la probabilidad de que esto ocurra.

La cantidad

$$\frac{d - \delta}{\sqrt{\frac{2S^2}{n}}} = \frac{d - 20}{6} = t_2$$

tiene una distribución de  $t$  con  $g=14$ . Pero es evidente que  $d \geq 10.5678$  si, y sólo si,

$$t_2 = \frac{d - 20}{6} \geq \frac{10.5678 - 20}{6} = -1.572$$

La probabilidad que se busca, por tanto, es la de que un valor de  $t$  con  $g=14$  sea mayor de -1.572. En la tabla  $t$ , la probabilidad de que un valor se encuentre fuera de los límites  $\pm 1.572$  es cerca de 0.14. De aquí que la probabilidad deseada sea

$$1 - \frac{0.14}{2} = 0.93$$

o sea, cerca de 93%. Por lo que es muy probable que en este experimento se encuentre una diferencia significativa.

Pero en la práctica, la cuestión que se plantea es el querer saber cuál es el tamaño mínimo de muestra que se debe tomar para poder detectar una diferencia mínima,  $\delta$ , verdadera, con una probabilidad  $P$  al nivel de significación  $\alpha$ .

Con arreglo a lo visto hasta el momento, la probabilidad aproximada,  $P$ , de obtener un resultado significativo se puede calcularla haciendo uso de la siguiente expresión

$$\delta = \sqrt{\frac{2\sigma^2}{n}}(t_1 + t_2)$$

donde  $\delta$  = diferencia verdadera que se desea estimar.

$\sigma$  = error típico verdadero por unidad

$r$  = número de repeticiones.

$t_1$  = valor significativo de  $t$  en la prueba de significación (hay que tener cuidado en distinguir entre prueba de una o de dos colas)

Como

$$P = 1 - \frac{P_2}{2}$$

donde  $P_2$  es la probabilidad correspondiente a  $t_2$  de la tabla. Los grados de libertad de  $t_1$  y  $t_2$  son los de la varianza del error.

De esta fórmula podemos despejar  $n$  con objeto de dar los tamaño de muestras  $n$  requeridas para una probabilidad dada  $P$  de obtener un resultado significativo. En esta forma, se tiene

$$n \geq 2 \left( \frac{\sigma}{\delta} \right)^2 [t_{(g); \alpha} + t_{(g); 2(1-P)}]^2$$

donde  $n$  = tamaño de muestra.

$\sigma$  = desviación típica verdadera por unidad

$\delta$  = diferencia verdadera mínima que se desea detectar.

$P$  = probabilidad de que  $\delta$  de significativo.

Esta inecuación se resuelve por iteraciones, pues los grados de libertad de ambas  $t$  dependen de  $n$ , por lo que tendremos que hacer varios tanteos antes de que encontremos que se cumpla la desigualdad.

Los valores de  $\delta$  y de  $\sigma$  suelen plantearse en proporciones, y aunque el verdadero valor de  $\sigma$  habitualmente sea desconocido la mayoría de las veces, es suficiente con conocer el cociente  $\sigma/\delta$ .

Nótese, así mismo que  $1-P = \beta$ .

### Ejemplo.-

Por estudios anteriores se sabe que el coeficiente de variación de las longitudes de alas de una especie de ave es aproximadamente del 6%. Se necesita estudiar 4 poblaciones ¿Cuántas medidas serán necesarias, de cada población, para tener la seguridad del 80% de detectar un 5% de diferencia entre dos de las cuatro medias a un nivel de significación del 1%?

Dado que se tiene cuatro grupos  $g=4(n-1)$ .

Aplicando la fórmula anterior, se puede resolver por medio de iteraciones. Para ello se empieza dándole un valor a  $n$ , que para comenzar puede ser, por ejemplo,  $n=20$  como un posible valor razonable. Entonces  $g=4(20-1)=76$ . Dado que  $CV=6\%$ ,

$$S = \frac{6\bar{X}}{100}$$

Se desea que la potencia de la prueba ( $\delta$ ) sea el 5% de la media, es decir que

$$\delta = \frac{5\bar{X}}{100}$$

Utilizando  $S$  como estima de  $\sigma$  de tiene que

$$\frac{\sigma}{\delta} = \frac{\frac{6\bar{X}}{100}}{\frac{5\bar{X}}{100}} = \frac{6}{5}$$

Por lo que

$$\begin{aligned} 20 &\geq 2 \left(\frac{6}{5}\right)^2 [t_{(76; 0.01)} + t_{(76; 2(1-0.80))}]^2 = \\ &= 2 \left(\frac{6}{5}\right)^2 (2.64+0.847)^2 = 2 \times 1.44 \times 12.16 = 35.0 \end{aligned}$$

Como es evidente que no se cumple la desigualdad ( $20 > 35$ ), se vuelve a resolver la inecuación, ahora con valor, por ejemplo,  $n=35$ , con cuyo valor  $g=4(35-1)=136$

$$35 \geq 2 (1.444) (2.61+0.845)^2 = 34.4$$

Ahora si se cumple la inecuación ( $35 > 34.4$ ), siendo muy cercanos los valores, lo que indica que se ha iterado (repetido) con estabilidad. Por lo que el resultado final es de  $n=35$ .

## Comparación de dos proporciones binomiales.-

En el Capítulo anterior se vio varios ejemplos en los que, haciendo uso del Teorema Central del Límite, se contrastaba si una proporción determinada era igual o no a una proporción teórica esperada por una hipótesis. Se realizaba esta aproximación a la distribución normal por tener un tamaño de muestra grande. Por ejemplo, se espera que al lanzar una moneda al aire salga cara con  $p=0.5$ , por lo que se puede hacer la experiencia 100 veces y comprobar si efectivamente la  $p$  es 0.5. Otro ejemplo es, se puede hacer cruces de híbridos para el color del pelo y esperar que en la descendencia haya 0.75 de una coloración y 0.25 de la otra.

En este Capítulo se va a comparar dos poblaciones de las que se han extraído muestras para medir la proporción de cierto fenómeno y se quiere saber si la proporción es la misma o no en ambas poblaciones. Por ejemplo, comparación de la tasa de incidencia de cierta enfermedad en dos regiones, comparación de la proporción de enfermos de cáncer en individuos no fumadores y en individuos fumadores, etc. Como se ve se trata de variables cualitativas y no cuantitativas como las que se exigen para la prueba  $t$  de contraste de medias.

Sean las dos muestra siguientes

$$X_{11}, X_{12}, \dots, X_{1n_1} \sim \text{Bin}(n_1, p_1)$$

$$X_{21}, X_{22}, \dots, X_{2n_2} \sim \text{Bin}(n_2, p_2)$$

La dos muestras son independientes y las probabilidades  $p_1$  y  $p_2$  se relacionan con la ocurrencia del mismo suceso bajo dos condiciones experimentales distintas. Para comparar las dos situaciones experimentales se usará el parámetro  $p_1 - p_2$ .

Los estimadores puntuales de  $p_1$  y  $p_2$  son

$$\hat{p}_1 = \frac{X_{11}, X_{12}, \dots, X_{1n_1}}{n_1} = \frac{X_1}{n_1}$$

$$\hat{p}_2 = \frac{X_{21}, X_{22}, \dots, X_{2n_2}}{n_2} = \frac{X_2}{n_2}$$

Puesto que ambas variables son binomiales se tiene

$$E(\hat{p}_1) = E \frac{X_1}{n_1} = \frac{n p_1}{n_1} = p_1$$

$$\sigma^2_{(\hat{p}_1)} = \sigma^2 \frac{X_1}{n_1} = \frac{1}{n_1^2} \sigma^2_{(X_1)} = \frac{p_1(1-p_1)}{n_1}$$

$$E(\hat{p}_2) = E \frac{X_2}{n_2} = \frac{n p_2}{n_2} = p_2$$

$$\sigma^2_{(\hat{p}_2)} = \sigma^2 \frac{X_2}{n_2} = \frac{1}{n_2^2} \sigma^2_{(X_2)} = \frac{p_2(1-p_2)}{n_2}$$

El Teorema Central del Límite asegura que para  $n_1$  y  $n_2$  grandes

$$\hat{p}_1 \sim N \left( p_1, \frac{p_1(1-p_1)}{n_1} \right)$$

$$\hat{p}_2 \sim N \left( p_2, \frac{p_2(1-p_2)}{n_2} \right)$$

Como tanto  $\hat{p}_1$  como  $\hat{p}_2$  son medias de variables aleatorias independientes, se tiene

$$\hat{p}_1 - \hat{p}_2 \sim N \left( p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \right)$$

Si se pretende probar las hipótesis

$$H_0 : p_1 - p_2 = 0$$

$$H_1 : p_1 - p_2 \neq 0$$

Hay que generar un estadístico de prueba y, tal como se hizo con la diferencia de medias, hay que notar que bajo la  $H_0$  (igualdad de proporciones) se tiene

$$\hat{p}_1 - \hat{p}_2 \sim N \left( 0, \frac{P(1-P)}{n_1} + \frac{P(1-P)}{n_2} \right)$$

en el que el único parámetro desconocido es  $P$  o proporción conjunta. La estima de este parámetro es la proporción ponderada del suceso en las dos poblaciones, es decir

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

Siendo, por tanto, el error típico de la diferencia de proporciones

$$S_{p_1-p_2} = \sqrt{\hat{p}(1-\hat{p}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

Y el estadístico que permitirá probar la hipótesis es una prueba  $Z$  tal que

$$Z_0 = \frac{\hat{p}_1 - \hat{p}_2}{S_{p_1-p_2}}$$

Si se quiere hacer por intervalo de confianza este es

$$L_i = \hat{p}_1 - \hat{p}_2 - Z_{(\alpha)} S_{p_1-p_2}$$

$$L^s = \hat{p}_1 - \hat{p}_2 + Z_{(\alpha)} S_{p_1-p_2}$$

Esta aproximación a la distribución  $Z$  de una distribución binomial, se puede realizar siempre que los tamaños de las muestras sean grandes. En el caso de que se tengan muestras pequeñas, se puede realizar este mismo contraste por medio de una de las pruebas para datos categóricos que se estudiarán en el Capítulo 20 dedicado a los **Datos Categóricos**.

### Ejemplo.-

Un sociólogo de cierto país desea comparar los porcentajes de analfabetismo en dos comunidades, una *rural* y otra *urbana*. Su hipótesis es de que el analfabetismo es mayor en las comunidades rurales que en las urbanas, para ello hace una encuesta a 326 individuos de la comunidad rural y a 381 de la comunidad urbana, encontrando 62 analfabetos en el campo 39 en la ciudad.

*campo:*      *ciudad:*

$$n_1 = 326 \quad n_2 = 381$$

$$X_1 = 62 \quad X_2 = 39$$

$$H_0 : p_1 \leq p_2$$

$$H_1 : p_1 > p_2$$

$$\hat{p}_1 = \frac{62}{326} = 0.190$$

$$\hat{p}_2 = \frac{39}{381} = 0.102$$

$$\hat{p} = \frac{326 \times 0.19 + 381 \times 0.102}{326 + 381} = 0.1426$$

$$S_{(p_1-p_2)} = \sqrt{0.1426 \times 0.8574 \left( \frac{1}{326} + \frac{1}{381} \right)} = 0.026379$$

$$Z_o = \frac{0.190 - 0.102}{0.026379} = 3.34^{**}$$

$$Z_{(0.01)} = 2.33$$

Se rechaza la  $H_0$  con un nivel de significación de  $\alpha=0.01$ . Luego se concluye que la hipótesis del sociólogo es cierta

Si se hace con el intervalo de confianza

$$L_i = 0.190 - 0.102 - 2.33 \times 0.02638 = 0.0265$$

$$L^s = 0.190 - 0.102 + 2.33 \times 0.02638 = 0.1495$$

No incluye al cero, por lo que, en este país, el porcentaje de analfabetismo, en el campo es mayor que en la ciudad, con un 99% de confianza.

## Bibliografía

- Afifi, A.A. y Clark, V.* 1984. COMPUTER-AIDED MULTIVARIATE ANALYSIS. Ed: Lifetime Learning Publications. Belmont (USA).
- Bisquerra Alzina, R.* 1989. INTRODUCCIÓN CONCEPTUAL AL ANÁLISIS MULTIVARIABLE. Ed: PPU. Barcelona (España).
- Cuadras, C.M.* 1981. MÉTODOS DE ANÁLISIS MULTIVARIANTE. Ed: EUNIBAR. Barcelona (España).
- Dagnelie, P.* 1982. ANALYSE STATISTIQUE À PLUSIEURS VARIABLES. Ed: Les Presses Agronomiques De Gembloux. Gembloux (Belgique).
- Dixon, J.D. y Manssey, F.J.* 1974. INTRODUCCIÓN AL ANÁLISIS ESTADÍSTICO. Ed. Del Castillo. Madrid.
- González López-Valcárcel, B.* 1991. ANÁLISIS MULTIVARIANTE: APLICACIÓN AL ÁMBITO SANITARIO. Ed: SG Editores. Barcelona (España).
- Infante Gil, S. y Zárate De Lara, G.P.* 1984. MÉTODOS ESTADÍSTICOS. Ed. TRILLAS. México.
- Judez Asensio, L.* 1989. TÉCNICAS DE ANÁLISIS DE DATOS MULTIDIMENSIONALES. Ed: MAPA. Madrid (España).
- Lebart, L., Morineau, A. Y Fénelon, J.P.* 1979. TRAITEMENT DES DONNÉES STATISTIQUES. Ed: Dunod. Paris (France).
- Lefebvre, J.* 1980. INTRODUCTION AUX ANALYSES STATISTIQUES MULTIDIMENSIONNELLES. Ed: Masson. Paris (France)
- Li, Ching Chun.* 1969. INTRODUCCIÓN A LA ESTADÍSTICA EXPERIMENTAL. Ed. Omega. Barcelona.
- Littel, R.C., Freund, R.J. and Spector, P.C.* 1991. SAS FOR LINEAR MODELS. SAS Institute Inc., Cary, NC, USA.
- Milton, J.S.* 1994. ESTADÍSTICA PARA BIOLOGÍA Y CIENCIAS DE LA SALUD. Ed. Interamericana-McGraw-Hill. México.
- Snedecor, G.W. y Cochran, W.G.* 1971. MÉTODOS ESTADÍSTICOS. Ed. C.E.C.S.A. México.
- Sokal, R.R. y Rohlf, F.J.* 1994. BIOMETRY. Ed. W.H.FREEMAN. San Francisco.
- Spiegel M.R.* 1990. ESTADÍSTICA. Ed. McGraw-Hill. Madrid.
- Srivastava, M.S. y Carter, E.M.* 1983. AN INTRODUCTION TO APPLIED MULTIVARIATE STATISTICS. Ed: Elsevier Science Publishing. New York (USA).
- Steel, R.* 1996. PRINCIPLES AND PROCEDURES OF STATISTICS. Ed. McGRAW-HILL Education. New York .
- SAS Institute Inc. 1990. SAS/STAT USER'S GUIDE. Volume 1 and 2. Cary, NC, USA.

## **CAPÍTULO 5**

### **Análisis de la Varianza y Análisis Multivariante de la Varianza**





# Análisis de la Varianza y Análisis Multivariante de la Varianza

### Generalización.-

Hasta ahora se han estudiado los diseños muestrales-experimentales con una y dos muestras. Las inferencias para estos dos tipos de diseños se realizan con las mismas técnicas analíticas, con la condición de que, en el caso del diseño experimental, los tratamientos que reciban las diferentes unidades experimentales lo sea de una manera completamente aleatoria; o, en el caso del diseño muestral, que el muestreo que se haga en cada una de las dos poblaciones sea, también, de una manera completamente aleatoria.

Pero, también se ha visto, que no todos los diseño muestrales o experimentales son completamente aleatorios (con *muestras independientes*), sino que hay diseños con *muestras emparejadas*.

Cuando se estudió el caso de muestras emparejadas se vio que respondía a un diseño experimental o muestral que incluía una *estratificación* de la población que tenía como consecuencia una información, y por lo tanto un análisis, más precisa que el muestreo o el diseño completamente aleatorio. Para esta estratificación, la población se subdivide en varias partes o estratos de una manera no aleatoria, esto es, es el experimentador el que decide cuáles son los diferentes estratos con arreglo a criterios de clasificación presentes en la población, y dentro de cada estrato se hace un diseño muestral o experimental completamente aleatorio.

Tanto si se trata de diseños experimentales como si se trata de diseños muestrales es muy importante, si se quiere realizar inferencias correctas, que se sepan ver estos estratos con objeto de tenerlos en cuenta en el análisis.

A partir de ahora se va a generalizar estas inferencias en dos aspectos

- 1) El número de poblaciones serán más de dos

- 2) El número de clasificaciones o sistemas de estratos, también será uno o más de uno.

También hay que aclarar que la presente exposición estará enfocada más desde la perspectiva de los diseños muestrales que desde la perspectiva de los diseños experimentales, si bien, como se ha dicho anteriormente, las técnicas de análisis son las mismas para ambos tipos de diseños

### **Ideas elementales sobre los Modelos.-**

Una práctica corriente en la Ciencia es tratar de explicar fenómenos naturales mediante *modelos*. Un modelo es una representación reducida de la realidad, es decir, de la complejidad de la realidad se modeliza la esencia de esta, la que se quiere explicar. Dicho de otro modo, un modelo es una interpretación de un hecho abstracto o inobservable expresado en términos de hechos no abstractos u observables. Por ejemplo, se puede decir que la *Tierra es como un globo* (esférica) y gira sobre un eje Norte-Sur. Que la Tierra es redonda es un hecho conjeturable y por lo tanto inferible pero no es un hecho observable (nadie ha visto la totalidad del planeta, ni siquiera los astronauta pueden verla en su totalidad) y para comprender ciertos hechos se utiliza un modelo no abstracto y por lo tanto observable, esto es, que la Tierra es como el globo que tenemos sobre la mesa. Este es un modelo verbal que permite comprender y explicar multitud de observaciones y hechos empíricos, y permite prever resultados de experiencias aún no realizadas, como puede ser la de navegar siempre hacia el Oeste para volver al lugar de partida, lo que confirmará la veracidad (o falsedad) del modelo.

Modelos más perfeccionados dependen de afirmaciones matemáticas que permiten explicar los hechos observables (los datos), por medio de hechos no observables (adelantemos como hecho no observable, el término de *efecto de un factor*). El aspecto más importante de un modelo matemático es que posee las características intrínsecas del problema real.

### **El Modelo lineal aditivo.-**

Cuando se ha afirmado, anteriormente, que este manual esta enfocado desde la perspectiva de los diseños muestrales, estamos haciendo hincapié en el hecho de que los diseños son, en principio, infinitos, y que ante un problema hay que estudiar cual es el modelo que nos provee de la información más pertinente con el mínimo de esfuerzo y dinero. La consecuencia de este modelo será la determinación del diseño (experimental o muestral) que se tiene que utilizar y el análisis adecuado para la precisión deseada.

La calidad de un análisis estadístico esta directamente relacionado con la adecuación del modelo que se asume para describir los datos. Un modelo debe representar adecuadamente la característica muestral de los datos y reflejar el problema biológico. Hay tres niveles de modelos.

- 1) Modelo *verdadero* que describa los datos perfectamente, sin variación residual o no explicada. El modelo verdadero nunca es exactamente conocido.
- 2) Un modelo *ideal* es una formulación del experimentador tan próximo al modelo verdadero como sea posible. El modelo ideal tendría que ser el modelo que se usara para el análisis, pero a menudo no se tiene suficiente información disponible como para aplicarlo en la práctica.
- 3) Modelo *operacional* es una versión simplificada del modelo ideal, y es el que el experimentador usara en el análisis. A menudo no hay acuerdo entre los investigadores de cual es el mejor modelo operacional a usar.

Para elaborar un modelo operacional correcto, el experimentador debe construir un modelo ideal. Dadas las limitaciones de los datos y los recursos bajo los cuales el experimentador debe funcionar, el modelo operacional puede elaborarse simplificando apropiadamente el modelo ideal. Si se realizan demasiadas simplificaciones o se imponen demasiados supuestos, entonces puede ocurrir que el análisis no sea válido. Siguiendo esta aproximación para construir un modelo, el experimentador debe conocer exactamente los supuestos que debe imponer, lo que implicará la mayor o menor calidad del modelo operacional. Esto quiere decir que es tan importante lo que se explicita en el modelo como lo que no se incluye.

Los modelos contendrá un grupo de *factores* que afectan aditivamente al valor de la variable.

Los factores pueden ser discretos o continuos y las variables son asignadas a estos porque, suponemos, pueden influir en los valores de las observaciones. Por ejemplo, en la producción de leche es conocida la influencia de la edad al parto, de la estación de parto, del potencial genético, y del número de días entre partos, por citar algunos.

Para una correcta elaboración del modelo se requiere que sean identificados la mayoría de los factores usuales.

Los factores discretos tienen clases o niveles tal como la edad al parto que puede tener cuatro niveles (es decir, de 20 a 24 meses, de 25 a 28 meses, de 29 a 32 meses, y de 33 o más meses). Por lo que un análisis puede proveer estimas de diferencias en producciones de leche de vacas de varias edades. Alternativamente, el efecto de la edad puede ser considerado como una covariable (ver Capítulo 17 dedicado al **Análisis de Covarianza**) con un efecto lineal y cuadrático sobre la producción de leche, y pueden estimarse los coeficientes de regresión relativos a edad y edad al cuadrado para producción de leche.

Algunos factores pueden tener un interés especial para los investigadores mientras que otros factores hay que incluirlos en el modelo con objeto de reducir la variabilidad residual. Por ejemplo, un investigador puede estar interesado en los efectos de varios niveles de dosis de la hormona del crecimiento en terneras, pero el modelo debe también incluir los efectos de la edad del animal, el sexo, la

localización, dieta, y raza. El último grupo de efectos son a menudo denominados factores *ruido*. Los factores ruido no pueden ser ignorados u omitidos del modelo porque esto puede alterar drásticamente la interpretación de los resultados del factor principal de interés.

En un estudio estadístico tradicional también se necesita distinguir entre factores de efectos fijos y de efectos aleatorios, pero esto se verá más adelante.

Los modelos lineales son adecuados en la mayoría de las circunstancias biológicas. Esto no implica que modelos no lineales no tengan importancia. Como se verá más adelante, relaciones no lineales pueden a menudo aproximarse por modelos lineales. Pero si un modelo no lineal da mejor modelo ideal que un modelo lineal, entonces debe ser utilizado el modelo no lineal.

Pero un modelo lineal aditivo implica no solo la expresión matemática o ecuación, sino el conjunto de supuestos, restricciones y limitaciones implícitas en dicho modelo.

Estos supuestos incluyen detalles que no son aparentes en la expresión matemática; por ejemplo, información de los datos o de la manera en que los datos se han tomado. ¿Fueron seleccionados aleatoriamente las unidades experimentales o fueron tomados a partir de mínimos estándar? ¿Se criaron en ambientes aleatorios o en ambientes especialmente elegidos? *etc.*

Este aspecto del modelo debe recoger las diferencias entre el modelo operacional y el modelo ideal y los posibles efectos de estas diferencias deben ser explicadas en el análisis. Tal comparación es frecuentemente subvalorada o ignorada, pero este aspecto del modelo contiene la información mas importante para calibrar la calidad del análisis.

El diseño muestral, el procedimiento estadístico y la estrategia del análisis de los datos se determinan sólo después de completado el modelo en todos sus aspectos.

Por ejemplo, se esta estudiando el peso en terneros entre los 190 y los 210 días, para ello se realizaron cuatro grupos de edad de intervalos de cinco días, y se asumió que este factor era de efectos fijos, por lo que le modelo podría se el siguiente.

$$X_{ij} = \mu + E_i + e_{ij}$$

donde

$X_{ij}$  es el peso de la  $j$ -ésima ternera dentro del  $i$ -ésimo grupo de edad.

$\mu$  es la media general

$E_i$  es el efecto de la  $i$ -ésima edad.

$e_{ij}$  es el efecto residual.

El segundo aspecto del modelo incluye lo siguiente:

1. Raza de los terneros es la misma.
2. Edad de la madre no influye en el peso a estos días.
3. Todos los terneros son del mismo sexo.

4. Todos los terneros son no parientes.
5. Todos los terneros fueron criados en el mismo ambiente y bajo el mismo régimen alimenticio.

Cada uno de estos estamentos deben soportar varios tipos de evidencias. Por ejemplo, sería necesario un estudio que mostrara que edad de hembras no influye en el peso a estos días. Si la evidencia es escasa o contradictoria entonces quizás deba incluirse en el modelo la edad de la madre.

El modelo lineal aditivo, por tanto, trata de explicar una observación como una suma de componentes tales como la media paramétrica más unos elementos de variación aleatorios que pueden ser asociados a diferentes efectos, fuentes de variación o factores (ver más adelante) más un *error* que incluirá tanto los *errores de medida* como los *errores de muestreo* y el *error experimental*.

Si tomamos cualquier libro de estadística veremos que vienen descritos muchos tipos de análisis de la varianza correspondientes a otros tantos diseños o modelos. Este manual se dedica a desarrollar los análisis de varianza como respuesta a diseños de muestreo, en el que la labor de diseño está en la creación del modelo lineal adecuado al problema planteado y a los datos o experiencia desarrollada. Es el mismo proceso del diseño de experimentos pero con experiencias que vienen obligadas en gran parte, como es el caso de los trabajos realizados con animales domésticos o con humanos.

Como veremos más adelante, cualquier modelo o diseño, por complejo que sea, se puede catalogar como uno de estos tres tipos

- 1) Los que incluyen sólo efectos jerárquicos.
- 2) Los que incluyen sólo efectos factoriales.
- 3) Y los que contengan efectos jerárquicos y factoriales.

Dándoseles el nombre de diseño o experimento *Jerárquico*, diseño o experimento *Factorial* y experimento *Compuesto*, respectivamente.

Un efecto está jerárquico (encajado, anidado) respecto a otro cuando está subordinado a él; por ejemplo, se tiene tres machos a prueba, para ello se toman seis hembras, dos para cada macho, en este caso el factor hembra está jerárquico al factor macho.

Unos efectos son factoriales si tienen igual categoría entre ellos, de manera que se consideran todas las combinaciones posibles entre las diferentes clases de cada uno de los factores; por ejemplo, se aplican tres tratamientos a dos razas, probando todas las razas en cada tratamiento y los tres tratamientos en cada raza, es decir, se tendrán presentes las seis combinaciones de raza y tratamiento. Aquí los efectos de raza y tratamiento son factoriales.

## Factor, tratamientos y efectos.-

Aunque ya se ha hablado de factores en el epígrafe anterior, en realidad el modelo lineal estudiado hasta el momento es, con mucho, el más simple que se puede presentar pues solo lleva la media paramétrica ( $\mu$ ) y el error experimental ( $\epsilon$ ), pero su generalización permitirá analizar situaciones de mayor complejidad. Comencemos estas generalizaciones.

Al principio de esta manual se hizo la afirmación de que la estadística sirve para cuantificar la incertidumbre de las inferencias basadas en el estudio de una variable aleatoria medidas en muestras. Esto es lo que se ha hecho en el epígrafe anterior.

Pero también se afirmó que la estadística permite medir las pequeñas, pero significativas, diferencias de algunos de los factores que propicia que los fenómenos naturales se nos presenten como variables. Esto es lo que se va a estudiar a partir de ahora, si bien se hizo una introducción anteriormente en el que se habló de los términos *tratamiento* y *efecto del tratamiento*.

Expliquemos lo que quiere decir el párrafo anterior. Cuando se estudió los errores dos epígrafes atrás se afirmó que el error de muestreo era el debido a las diferentes causas que intervienen en la medida final de la variable. Se puso como ejemplo la altura de las personas y se dijo que todas las personas adultas del mismo sexo y misma población medirían lo mismo si tuvieran el mismo fondo genético (sus padres midieran lo mismo), se hubieran criado en el misma familia, en el mismo barrio, la misma ciudad,..., hubieran padecido las mismas enfermedades, hubieran comido lo mismo, hubieran sufrido los mismos accidentes, hubieran tenido las mismas vivencias, ..., en definitiva, si fueran la misma persona. Como eso no es así, las personas tienen diferentes alturas. Ahora bien, además de cuantificar la incertidumbre que se tiene al inferir cual es la altura esperada de esa población de personas si se toma una muestra de diez personas, cosa que se ha hecho en el epígrafe anterior, también se podría, de entre esa multitud de causas que intervienen en la estatura final de las personas, detectar diferencias entre algunas de dichas causas. Por ejemplo ¿Diferencia en las dietas propicia diferencia en la altura? se puede detectar, si las hubiera, estas diferencias si se toman 10 niños y se le da un tipo de dieta y otros 10 niños y se le da otro tipo de dieta y al final del crecimiento se comprueba cual grupo tiene una mayor estatura media. Como esta experiencia sería muy larga y, probablemente, legalmente imposible, imaginemos el mismo ejemplo para engorde de cerdos.

Efectivamente, el análisis de varianza se plantea como un análisis para detectar efectos de diferentes tratamientos o diferentes niveles de un factor. Para entender estos términos se va a utilizar el mismo símil de *Li*.

Para celebrar su 84 cumpleaños, la abuela decide distribuir 84 **euros** entre sus 12 nietos. Siendo imparcial y queriéndolos a todos lo mismo, la abuela divide los 84 **euros** entre los 12 nietos por partes iguales ( $84=12*7$ ). El abuelo advierte que 4 de los nietos son licenciados universitarios, 4 están estudiando en la Universidad y los cuatro restantes están todavía en el *ESO*. El no cree que todos deban tener la misma cantidad de dinero para sus gastos y sugiere que los mayores reciban más que los jóvenes. Cortésmente, cada uno de los 4 estudiantes de *ESO* le ceden 2 **euros** al

grupo de los licenciados, y cada uno de los cuatro estudiantes universitarios le ceden 1 **euro** a los mayores. La cantidad total de dinero entre los 12 nietos sigue siendo 84 **euros**, pero ahora los licenciados tienen un total de 40 **euros**, poseyendo cada uno de ellos 10 **euros** en lugar de los 7 **euros** iniciales, mientras que los estudiantes universitarios tienen cada uno 6 **euros** y los estudiantes de *ESO* tiene cada uno 5 **euros**. Después de la redistribución, deciden jugar a las cartas en tres grupos separados, los licenciados por un lado, los estudiantes universitarios por el otro y los de *ESO* en el tercer grupo. Todos aceptan la regla de que la unidad de apuesta sean monedas de 5 centavos. Al final de la tarde, cada muchacho cuenta lo que ha ganado o perdido en el juego y cuánto le queda de los 7 **euros** recibidos al principio. En la siguiente tabla se detalla lo ocurrido a cada uno de los muchachos.

	<i>Cantidad Original</i> $\bar{X} ..$	<i>Aportación</i> $\tau$	<i>Comienzo Juego</i> $\bar{X} i.$	<i>Ganancia juego</i> $\epsilon_{ij}$	<i>Final juego</i> $X_{ij}$	<i>Desviación total</i> $X_{ij} - \bar{X} ..$
<i>ESO</i>	7	-2	5	+1	6	-1
	7	-2	5	-3	2	-5
	7	-2	5	+4	9	+2
	7	-2	5	-2	3	-4
<i>Universitarios</i>	7	-1	6	-2	4	-3
	7	-1	6	+4	10	+3
	7	-1	6	-2	4	-3
	7	-1	6	0	6	-1
<i>Licenciados</i>	7	+3	10	-1	9	+2
	7	+3	10	+2	12	+5
	7	+3	10	-2	8	+1
	7	+3	10	+1	11	+4
<i>Total</i>	84	0	84	0,0,0	84	0

Obsérvese que las pérdidas y ganancias del juego es cuestión de cada grupo; lógicamente, en cada grupo deben compensarse las pérdidas o ganancias entre los cuatro jugadores.

Poco después de concluida la partida llega el abogado de la familia para felicitar a la abuela por su cumpleaños. Entonces se entera del obsequio de la abuela y de la partida de cartas pero no conoce los detalles. Sabe el dinero que ahora tiene cada muchacho: para el grupo de los jóvenes, 6, 2, 9, 3; para el grupo universitario, 4, 10, 4, 6; para el grupo de licenciados, 9, 12, 8, 11, respectivamente. En otras palabras, todo lo que conoce son los resultados del proceso registrados en la columna 5 de la tabla anterior. Estas son sus observaciones o datos. ¿Puede el abogado, que nunca ha estudiado estadística, reconstruir, a partir de tan sólo sus observaciones, la serie completa de los hechos ocurridos en este proceso, es decir, la tabla anterior? Podrá si se le revelan dos condiciones generales.

- 1) En cada mesa de juego, al iniciarse la partida, cada jugador tenía la misma cantidad de dinero.



2) Originalmente, cada muchacho recibió de la abuela la misma cantidad.

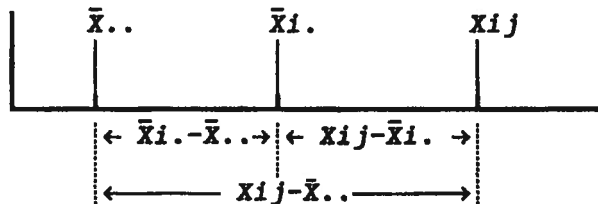
Informado de estas condiciones, el abogado puede reconstruir fácilmente toda la serie de hechos registrados en la tabla de la página anterior. Puesto que la cantidad total de dinero es 84 **euros**, en un principio cada muchacho debía tener 7 **euros**, que es la información de la columna 1. Y puesto que la cantidad total del grupo de estudiantes de *ESO* es 20 **euros**, al comienzo de la partida cada estudiante de *ESO* debía tener 5 **euros**, información de la columna 3. El primer estudiante de *ESO* termina con 6 **euros**; esto quiere decir que ha ganado 1 **euro** en la partida de cartas. Pero antes de comenzar la partida cedió 2 **euros**, de manera que el resultado neto es que todavía le falta un dólar para los 7 **euros** originales, y así sucesivamente.

El investigador (estadístico) está en la misma situación que el abogado de la familia. Todo experimento o estudio científico se basa en una serie de observaciones que son el resultado final de la acción de diversos factores. El trabajo estadístico es reconstruir, a partir de las observaciones, algo semejante a lo mostrado en la tabla de la página anterior y tratar de interpretar las observaciones desde ese punto de vista. Se puede examinar nuevamente dicha tabla explicando algunas columnas en lenguaje estadístico. Los números de la columna 5,  $X_{ij}$ , son las observaciones o datos. El número 7 de la columna 1 recibe el nombre de media general,  $\bar{X}_{..}$ , y los números de la columna 3 son las medias de los  $i$  grupos ( $\bar{X}_i$ ). La relación que existe entre estos valores puede resumirse en la siguiente expresión

$$(X_{ij} - \bar{X}_{..}) = (\bar{X}_i - \bar{X}_{..}) + (X_{ij} - \bar{X}_i)$$

es decir, la desviación total de un valor (lo que el individuo tiene al final con respecto a lo que tendría al principio o media general) es igual a la desviación del grupo en que está (media del grupo menos media general) más la desviación debida a la variación individual (lo que el individuo tiene al final con respecto a lo que tendría por el grupo al que pertenece).

Esta relación es fácilmente visualizable si se representa gráficamente de la siguiente manera



De esta última igualdad se obtiene

$$X_{ij} = \bar{X}_{..} + (\bar{X}_i - \bar{X}_{..}) + (X_{ij} - \bar{X}_i)$$

Por ejemplo, para el segundo individuo:  $2 = 7 + (5 - 7) + (2 - 5)$ ; y para el último individuo:  $11 = 7 + (10 - 7) + (11 - 10)$ .

Se puede examinar nuevamente la tabla para explicar, ahora, las otras columnas. Los números de la columna 2 son los *efectos* de los *i-ésimos* tratamientos, o efectos de los *i niveles* del factor; en este caso particular, podemos llamarlos efectos de los tres niveles del factor *edad* o efecto de los tratamientos *ESO, universitario y licenciado*. La ganancia o pérdida de la columna 4 se llama *error* o variación aleatoria de una observación individual o efecto del individuo. Finalmente, la última columna es la *desviación total*.

En el *error* de la columna 4 hay que incluir los dos errores anteriormente descritos: el de *medida* (que no se cuentan bien los dólares que tiene cada uno al final) y el de *muestreo* (que en la partida de cartas a algunos le cayeran con más frecuencia las cartas buenas y a otros las malas). Y además hay que incluir el otro tipo de error anteriormente citado pero no descrito, este es el *error experimental*. El error experimental se puede definir como la falta de uniformidad en la realización de un experimento, es decir, que el experimentador involuntariamente favorezca a unas unidades experimentales. En el ejemplo del cumpleaños de la abuela, el error experimental sería que uno (o algunos) hicieran trampa en el juego de cartas, con lo que las ganancias de algunos no serían aleatorias y, por tanto, no se corresponderían con las del grupo al que pertenece.

Una vez definidos los términos, podemos fijarnos en las estrechas y lógicas relaciones que hay entre las seis columnas de números. Por ejemplo, el segundo muchacho de *ESO* contribuyó, en un principio, con 2 **euros** (*efecto de la edad*) y después perdió en el juego otros 3 **euros** (*variabilidad individual* o error); por lo tanto pierde en total 5 **euros**. Así pues, la columna 6 es igual a la columna 2 más la columna 4. El efecto de la edad es  $\tau_i = \bar{X}_{i.} - \bar{X}_{..}$  y el del error es  $\varepsilon_{ij} = X_{ij} - \bar{X}_{i.}$ , por lo que la relación última se puede expresar, también, de la siguiente manera

$$X_{ij} = \bar{X}_{..} + (\bar{X}_{i.} - \bar{X}_{..}) + (X_{ij} - \bar{X}_{i.})$$

$$X_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

Hablando genéricamente, cada valor observado  $X_{ij}$  puede descomponerse en tres partes distintas; la primera parte,  $\mu$  es común a todas las observaciones; la segunda parte  $\tau_i$  es común a todos los miembros de cada grupo pero varía de un grupo a otro; y la tercera parte,  $\varepsilon_{ij}$ , varía de un individuo a otro.

En estadística, los tres grupos (en general,  $t$  tratamientos o  $t$  niveles del factor) de observaciones se interpretan como tres muestras independientes de una población grande de distribución normal cuya media general es  $\mu$ . Además se supone que esta población está formada por tres subpoblaciones, con la misma distribución (misma forma), pero con los parámetros de posición desplazados, siendo los valores medios respectivos:  $\mu_1=(\mu+\tau_1)$ ,  $\mu_2=(\mu+\tau_2)$ ,  $\mu_3=(\mu+\tau_3)$ . Por lo tanto, las muestras se han extraído de distintas subpoblaciones. En cada una de ellas, la desviación de un miembro respecto de la media de la subpoblación es  $\varepsilon_{ij}=X_{ij}-\mu_i$ . En consecuencia, cualquier miembro de la población puede expresarse en forma de una suma de distintas partes

$$X_{ij} = \mu_i + \varepsilon_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

que es la misma expresión anteriormente expuesta.

Este es otro modelo lineal, esencialmente igual al mostrado al principio, aumentándolo con la introducción de los efectos de los diferentes niveles de un factor o tratamientos.

El ejemplo del cumpleaños de la abuela conviene recordarlo para una mejor comprensión de los modelos que se verán a partir de ahora que, si bien serán más complicados, en esencia se podrán interpretar como este ejemplo.

### **Planteamiento del Análisis de la Varianza.-**

Como se ha dicho en el ejemplo del cumpleaños de la abuela, en la investigación científica o técnica nunca se va a conocer el proceso completo, sino que la situación del investigador es la del abogado de la familia que llega al final de la fiesta, vamos por tanto a plantear el problema tal como se presenta y comprobar si se puede explicar o reconstruir el proceso.

En cualquier campo de la investigación es frecuente encontrarse con problemas que tratan de analizar alternativas. Una dieta puede administrarse en distintas dosis (por ejemplo, tres dietas) e interesa saber si los efectos de las diferentes dietas sobre el rendimiento cárnico o lechero de la ganadería son diferentes. O bien son dietas para estudiar el crecimiento en niños. O se están probando tres antibióticos midiendo la disminución de linfocitos en sangre, etc. Estos son problemas que pueden resolverse con el ANOVA.

Cuando se mide una variable, como puede ser las producciones de una misma raza, el aumento de peso de niños, la cantidad de linfocitos en sangre, etc., el primer hecho que se pone de manifiesto es la variabilidad de dichas medidas. Centrándonos en el ejemplo de la producción ganadera; en el crecimiento y fisiología de los individuos influyen multitud de causas o efectos (*factores*), como pueden ser la constitución genética, cantidad de alimento ingerido, calidad del alimento ingerido, manejo, tipo de explotación, etc., actuando unos en un sentido, otros en otro sentido, y con mayor o menor intensidad. El resultado de todas estas acciones es, como se ha dicho, la existencia de diferencias en la producción. En principio, la variabilidad que aparece entre los animales sometidos a unos niveles del factor (dosis de pienso) y a otros, podrá atribuirse tanto a la dieta como a los restantes factores; por esta razón podría parecer que el procedimiento adecuado sería el mantener constantes todos los factores excepto el que se halla bajo estudio, analizando, entonces la respuesta a cada dosis. Esto no es posible, no sólo por el tipo y número de factores que entran en juego, sino por la existencia de causas desconocidas que pueden influir en la variabilidad. No pudiendo plantearse de esta manera, se podría plantear el problema de la siguiente manera: si la producción presenta una variabilidad causada por multitud de factores, teóricamente, es posible dividir esta variabilidad en dos partes, la originada por el factor estudiado, la dieta, y la variación producida por los restantes factores que entran en juego, sean conocidos o no, sean controlables o no, esto es lo que se conoce como *error* que engloba el *error de medida*, el *error de muestreo* y el *error experimental*. Una vez medidas estas dos variaciones, se puede hacer el contraste de si la variabilidad

*debida al factor* es, estadísticamente, mayor que la *debida al error*, lo que indicará si los efectos de los diferentes niveles del factor, o de los diferentes tratamientos, son iguales o no.

Todo esto no es más que el establecimiento de un modelo simple de comportamiento que se puede formalizar matemáticamente como un modelo lineal del tipo ya estudiado. El factor se presenta con  $i$  niveles o tratamientos y dentro de cada nivel se miden una serie de observaciones o unidades experimentales, así  $X_{ij}$  es la observación  $j$ -ésima del  $i$ -ésimo tratamiento. Este valor es descompuesto en dos partes  $\mu_i$  y  $\varepsilon_{ij}$ .  $\mu_i$  es la parte de  $X_{ij}$  debida al efecto del nivel  $i$ -ésimo que será común a todos los elementos sometidos a ese nivel del factor;  $\varepsilon_{ij}$  es la variación causada por todos los factores no estudiados, *error*, que variará de unas unidades a otras, dentro del mismo tratamiento. Las acciones de estas dos fuentes de variación se suponen lineales y aditivas, de manera que el modelo matemático sería

$$X_{ij} = \mu_i + \varepsilon_{ij}$$

Pero, como se ha visto con el ejemplo del cumpleaños de la abuela,  $\mu_i$  se puede descomponer en dos partes, una común a todas las observaciones que es  $\mu$  y otra privativa de cada nivel del factor o tratamiento que se ha simbolizado como  $\tau_i$ , siendo el modelo lineal

$$X_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

### Ejemplo.-

Para comprender la estima de los efectos se va a utilizar un ejemplo de números simples. Supóngase que se tienen 12 animales sometidos a tres tratamientos (4 animales por tratamiento) cuyas producciones totales, carácter medido con una determinada variable, y sus medias son las siguientes

$T_1$	$T_2$	$T_3$	
$X_{11} = 5$	$X_{21} = 6$	$X_{31} = 1$	
$X_{12} = 1$	$X_{22} = 2$	$X_{32} = 3$	
$X_{13} = 4$	$X_{23} = 7$	$X_{33} = 0$	
$X_{14} = 2$	$X_{24} = 1$	$X_{34} = 4$	
$X_{1.} = 12$	$X_{2.} = 16$	$X_{3.} = 8$	$X_{..} = 36$
$\bar{X}_{1.} = 3$	$\bar{X}_{2.} = 4$	$\bar{X}_{3.} = 2$	$\bar{X}_{..} = 3$
$\tau_1 = 0$	$\tau_2 = 1$	$\tau_3 = -1$	

En la producción de estos animales, actúan una serie de causas que se pueden dividir en tres

- 1) Un efecto común a todos debido a la producción media propia de la población ( $\mu$ ) que se estimará mediante  $\bar{X}_{..}$  (lo que la abuela le regala a cada nieto).
- 2) El efecto debido al tratamiento ( $\tau_i$ ) aplicado a cada *i-ésimo* grupo y que es un efecto controlado por el experimentador (el nuevo reparto de dinero que impone el abuelo con arreglo al factor edad de los nietos).
- 3) El efecto debido a la individualidad del animal y su situación ( $\varepsilon_{ij}$ ) (las partidas de cartas de los nietos), es decir, al resto de las infinitas causas que actúan sobre su producción y que se manifiestan como una causa única aleatoria. De estos factores, unos son conocidos y otros no; dentro de los conocidos podría haber unos que pueden ser medidos o controlados y otros que, pudiéndose medir, no se miden por alguna razón, fundamentalmente económica. Este efecto también incluye todo lo que no se controla como son los *errores de medida*, *errores de experimentación* y *errores de muestreo*, y su conjunto constituirá el llamado error o varianza residual.

Por lo tanto, el modelo es

$$X_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

Los valores de  $\mu$ ,  $\tau$  y  $\varepsilon$ , en la práctica, son estimados a partir de las  $X_{ij}$ , que son los números de los que se disponen. Sus valores reales son incognoscibles o inobservables, tan sólo se podrán obtener una estima de ellos mediante muestras.

La estima de la media general es:  $\mu = \bar{X}_{..} = 3$ .

Con arreglo al modelo, puesto que  $\mu_i = \mu + T_i$ , la estima de los efectos de los tratamientos se hace restando a la media de cada tratamiento la media general, de manera que, por ejemplo, para el tercer tratamiento  $t_3 = 2-3 = -1$ , es decir, ha disminuido en una unidad la media general, o lo que es lo mismo, ha disminuido la producción. El primer tratamiento ni aumenta ni disminuye la producción;  $t_1 = 3-3 = 0$ . Y el segundo tratamiento tiene como efecto aumentar la producción  $t_2 = 4-3 = 1$ . Como se observa, la suma de los efectos es igual a cero.

Estos valores de  $t_1$ ,  $t_2$  y  $t_3$  son completamente relativos pues bastaría cambiar uno de los tratamientos para que variaran los valores de  $t$ . Por lo que el valor de un efecto no sólo depende del valor intrínseco, sino que también depende de los otros tratamientos con que se compara.

Puesto  $X_{ij} = \mu_i + e_{ij}$ , para la estima de los efectos de los errores el procedimiento es semejante al empleado para la estima de los efectos de los tratamientos, es decir, se hace restando a cada valor la media del grupo de tratamiento al que pertenece. Por ejemplo, el animal  $X_{11}$  debiera haber producido 3 por pertenecer al tratamiento  $T_1$ , pero por causas desconocidas, este 3 se ha convertido en 5, por lo que las influencias de los factores no controlados han aumentado la producción en  $e_{11} = 5-3-0 = 2$  unidades,

mientras que para el individuo  $X_{14}$ , por ejemplo, estas causas desconocidas han actuado desfavorablemente, con lo que la estima de  $e_{14} = 2-3-0 = -1$ .

Estos valores que se obtienen para la estima de  $\epsilon$ , igual que las estimas de  $\tau$ , son también relativos, es decir, si los tres animales que acompañan a uno determinado fueran otros, su valor  $e_{ij}$  sería diferente. Si los nuevos son muchos peores que los anteriores, esa  $e_{ij}$  subirá, y si son peores bajará. Pero su suma siempre será, al igual que los efectos, cero.

Como las unidades experimentales utilizadas en una experiencia es una muestra aleatoria de las infinitas unidades que constituye la población, el error, tal como se dedujo en el modelo más simple, se distribuye como una normal de media cero y varianza la de la población, es decir

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

Como se ve, esta descomposición es la misma que tuvo que hacer el abogado de la familia en el cumpleaños de la abuela. La descomposición de los efectos de los doce individuos es la siguiente

			$\mu$		$t$		$e$		
$T_1$	<i>Animal</i>	$X_{11}$	5	=	3	+	0	+	2
		$X_{12}$	1	=	3	+	0	-	2
		$X_{13}$	4	=	3	+	0	+	1
		$X_{14}$	2	=	3	+	0	-	1
$T_2$	<i>Animal</i>	$X_{21}$	6	=	3	+	1	+	2
		$X_{22}$	2	=	3	+	1	-	2
		$X_{23}$	7	=	3	+	1	+	3
		$X_{24}$	1	=	3	+	1	-	3
$T_3$	<i>Animal</i>	$X_{31}$	1	=	3	-	1	-	1
		$X_{32}$	3	=	3	-	1	+	1
		$X_{33}$	0	=	3	-	1	-	2
		$X_{34}$	4	=	3	-	1	+	2

Esta tabla se ha elaborado a partir de los datos y es equivalente a la tabla del cumpleaños de la abuela vista anteriormente. Esta descomposición no se realiza normalmente, su exposición es sólo para aclarar el concepto de *efecto* como componente de la variable medida.

**Comparación de Tratamientos.-**

De la simple observación de los datos y de las medias del ejemplo anterior, y sin profundizar en su estudio, se podría deducir que el mejor tratamiento es el  $T_2$  pues es el que tiene efecto +1 y, por lo tanto, ha propiciado una mayor producción, mientras que el peor tratamiento es el  $T_3$  pues el que tiene efecto -1 y, por lo tanto, ha propiciado una menor producción.

Sin embargo, en el grupo del segundo tratamiento hay individuos con un valor igual o inferior al del peor tratamiento (el tres). Y observando la variación que se existe entre individuos de un mismo tratamiento se comprueba que esta variabilidad es mayor que la que se observa entre tratamientos; por ejemplo, la diferencia entre el mayor y el menor de cada tratamiento es de 4, 5 y 3, respectivamente, mientras que la diferencia entre tratamientos es de sólo 2.

Por lo que, a simple vista, no se puede concluir que el segundo tratamiento es significativamente mejor que el tercer tratamiento.

### Ejemplo.-

Para que sirva como elemento de discusión, supóngase otro ejemplo en el que las medias de los tratamientos, y por tanto los efectos, son las mismas que las del ejemplo anterior, pero las producciones de cada animal son diferentes. Estos datos son

$T_1$	$T_2$	$T_3$	
$X_{11} = 3$	$X_{21} = 4$	$X_{31} = 2$	
$X_{12} = 3$	$X_{22} = 3$	$X_{32} = 2$	
$X_{13} = 3$	$X_{23} = 5$	$X_{33} = 1$	
$X_{14} = 3$	$X_{24} = 4$	$X_{34} = 3$	
$X_{1.} = 12$	$X_{2.} = 16$	$X_{3.} = 8$	$X_{..} = 36$
$\bar{X}_1 = 3$	$\bar{X}_2 = 4$	$\bar{X}_3 = 2$	$\bar{X}_{..} = 3$
$\tau_1 = 0$	$\tau_2 = 1$	$\tau_3 = -1$	

Puesto que las medias de los tres tratamientos son iguales a los del primer ejemplo, se podría también afirmar que el  $T_2$  es el mejor pues su efecto es de +1, y el  $T_3$  el peor. Pero ¿En cuál de los dos ejemplos se estaría más seguros de la superioridad del  $T_2$  o de la inferioridad del  $T_3$ ? Parece evidente que se tendrá más seguridad en este último ejemplo, pues los cuatro individuos del mejor tratamiento ( $T_2$ ) tienen una medida mayor que los cuatro individuos del peor tratamientos ( $T_3$ ) y, también, tienen una medida mayor que los cuatro individuos del tratamiento intermedio ( $T_1$ ), excepto del valor tres que es igual. Mientras que en el ejemplo anterior, las medidas de los individuos del tratamiento con un efecto mayor ( $T_2$ ) se solapan con las medidas de los individuos de los tratamientos con efectos inferiores, de manera que solo dos individuos tienen medidas mayores que los individuos del tratamiento uno y tres, y los otros dos individuos podrían pertenecer al tratamiento uno o tres. Es decir, en el ejemplo anterior hay diferencias de cinco entre dos individuos de un mismo tratamiento, mientras que en el presente ejemplo el máximo de diferencia entre dos individuos del mismo tratamiento es de dos. Dicho de otra manera, en el ejemplo anterior, la variabilidad dentro de tratamiento (error) es muy superior a la variabilidad debida a los tratamientos, mientras que en el presente ejemplo ocurre todo lo contrario, por lo que en el primer ejemplo no se podría afirmar con seguridad que existe diferencia entre efectos de los tres tratamientos, mientras que en el presente ejemplo, siendo los efectos los mismos, si se puede afirmar con más seguridad que estas

diferencias son significativas. Más adelante se hará la prueba estadística de estos ejemplos, en las que se confirmarán estas apreciaciones intuitivas. Aunque en el epígrafe *Comparación de las medias de dos muestras aleatorias independientes* y siguientes se ha hecho la prueba estadística para un ejemplo como este pero con solo dos tratamientos.

Este modelo implica varias consecuencias que es necesario considerar.

- 1) La variable  $X_{ij}$  es una variable aleatoria normal con media  $\mu$  y varianza  $\sigma^2$ .
- 2) Es claro que deba suponerse que las  $\varepsilon_{ij}$  pueden tomar tanto valores positivos como negativos, puesto que suponer sólo valores positivos (o sólo negativos) equivaldría a suponer que en todas las medidas de la variable se obtienen lecturas mayores (o menores) que la verdadera. Por lo que es evidente, además, que  $\varepsilon_{ij}$  debe ser una variable aleatoria del tipo estudiado en el modelo más simple, por lo que el modelo para las  $\varepsilon_{ij}$  es Normal con media cero y varianza la de la variable.

$$X_{ij} \sim N(\mu, \sigma^2)$$

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

- 3) Las  $t$  muestras son muestras aleatorias e independientes con medias  $\mu_i$  y varianza la de la variable.

$$X_i \sim N(\mu_i, \sigma^2)$$

- 4) Los efectos de los tratamientos son una muestra de una distribución normal de media cero y varianza  $\sigma_T^2$ .

Simbolizando estos supuestos, se tiene

$$E(\varepsilon_{ij}) = 0; \sigma^2(\varepsilon_{ij}) = \sigma^2$$

$$E(\tau_i) = 0; \sigma^2(\tau_i) = \sigma_T^2$$

$$E(X_i) = E(\mu_i + \varepsilon_{ij}) = \mu_i + E(\varepsilon_{ij}) = \mu_i$$

$$\sigma^2(X_i) = \sigma^2(\mu_i + \varepsilon_{ij}) = \sigma^2(\varepsilon_{ij}) = \sigma^2$$

$$E(X_{ij}) = E(\mu + \tau_i + \varepsilon_{ij}) = \mu + E(\tau_i) + E(\varepsilon_{ij}) = \mu$$

$$\sigma^2(X_{ij}) = \sigma^2(\mu + \tau_i + \varepsilon_{ij}) = \sigma^2(\tau_i) + \sigma^2(\varepsilon_{ij}) = \sigma_T^2 + \sigma^2$$

Hay que aclarar que el cuarto supuesto no siempre es así. Este es para el caso que el factor sea de *efectos aleatorios*, es decir, para el caso en el que los diferentes niveles del factor o los diferentes tratamientos sean una muestra aleatoria de los muchos niveles del factor que existen y nos interesan. Para el caso de que sea un factor de *efectos fijos* no existe tal varianza pues los niveles son los únicos que existen, si se repite el experimento, siempre serán los mismos niveles. Mientras que si se repite



un experimento de efectos aleatorios, los niveles, al ser aleatorios, muy probablemente, serán otros. Por ejemplo, si se está probando tres antibióticos es un factor de efectos fijos con tres niveles, cada vez que se repita el experimento será con los mismos antibióticos, por lo tanto no habrá varianza. Por contra, si se está estudiando la influencia de la familia en la aparición de ciertas enfermedades, las familias serán los niveles de un factor de efectos aleatorios, si se tiene que repetir el experimento muy probablemente los será con otras familias.

El que, en este modelo, un factor sea de efectos fijos o de efectos aleatorios, no tiene más repercusión que a la hora de estimar parámetros, esta claro que no tiene sentido estimar la varianza de un factor de efectos fijos, pues esta no existe. Sin embargo, en modelos con más factores tiene otras repercusiones más difíciles de apreciar, por lo que estos dos tipos de efectos se estudiará más detenidamente unas páginas más adelante. Además de volverlo a estudiar más adelante.

### **Análisis de varianza de una Vía o del Diseño Completamente Aleatorio.-**

Esto se corresponde con el análisis de varianza de una vía o de clasificación única o de una sola fuente de variación o de un único factor, es decir, el análisis de varianza que se usa cuando las diferentes unidades experimentales o individuos estadísticos se han clasificado teniendo en cuenta un único criterio o se le ha aplicado los diferentes niveles de un único factor o tratamientos de una manera completamente aleatoria, tal como lo estudiado en el epígrafe anterior titulado *Planteamiento del análisis de la varianza*.

### **Modelo del análisis de Varianza de una Vía.-**

Este será el caso más sencillo de análisis de la Varianza y responderá al siguiente modelo lineal

$$X_{ij} = \mu + T_i + e_{ij}$$

donde  $\mu$  es la media general,  $T_i$  es el efecto del tratamiento  $i$ -ésimo y  $e_{ij}$  es el efecto del individuo  $j$  del tratamiento  $i$ .

### **Descomposición de la Varianza en un Modelo de una Vía.-**

La desviación de una observación cualquiera, respecto de la media general, con arreglo al modelo presentado, es el resultado de dos componentes

$$X_{ij} - \bar{X}_{..} = (\bar{X}_i - \bar{X}_{..}) + (X_{ij} - \bar{X}_i)$$

Siendo  $(\bar{X}_i - \bar{X}_{..})$  la desviación debida al efecto del tratamiento y  $(X_{ij} - \bar{X}_i)$  la desviación debida a la variabilidad individual o error.

Elevando al cuadrado tendremos

$$\begin{aligned}(X_{ij} - \bar{X}_{..})^2 &= [(\bar{X}_i - \bar{X}_{..}) + (X_{ij} - \bar{X}_i)]^2 \\ &= (\bar{X}_i - \bar{X}_{..})^2 + (X_{ij} - \bar{X}_i)^2 + 2(\bar{X}_i - \bar{X}_{..})(X_{ij} - \bar{X}_i)\end{aligned}$$

Si esto es cierto para una  $X_{ij}$  lo será para todas y cada una de las  $X_{ij}$ . Y también, teniendo en cuenta que al sumar para  $j$  un término en el que no existe  $j$  lo que hacemos es repetir dicho término tantas veces como unidades indique el valor de esta letra subíndice. La expresión de la suma de todas las diferencias para  $i, j$  sería

$$\sum_{ij} (X_{ij} - \bar{X}_{..})^2 = \sum_{ij} (\bar{X}_i - \bar{X}_{..})^2 + \sum_{ij} (X_{ij} - \bar{X}_i)^2 + 2 \sum_{ij} (\bar{X}_i - \bar{X}_{..})(X_{ij} - \bar{X}_i)$$

El último sumando tenemos que es igual a

$$\begin{aligned}\sum_{ij} (\bar{X}_i - \bar{X}_{..})(X_{ij} - \bar{X}_i) &= \sum_i (\bar{X}_i - \bar{X}_{..}) \sum_j (X_{ij} - \bar{X}_i) = \sum_i (\bar{X}_i - \bar{X}_{..})(\sum_j X_{ij} - \sum_j \bar{X}_i) = \\ &= \sum_i (\bar{X}_i - \bar{X}_{..})(\sum_j X_{ij} - \sum_j X_{ij}) = 0\end{aligned}$$

Luego el doble producto es cero.

Con lo que se llega a la conclusión ya conocida

$$\begin{aligned}\sum_{ij} (X_{ij} - \bar{X}_{..})^2 &= \sum_{ij} (\bar{X}_i - \bar{X}_{..})^2 + \sum_{ij} (X_{ij} - \bar{X}_i)^2 \\ SC_{total} &= SC_{Tratamientos} + SC_{Error}\end{aligned}$$

Estas son las fórmulas teóricas, las fórmulas prácticas de los diferentes términos de esta ecuación son las siguientes

$$SC_{Trata} \quad \sum_{ij} (\bar{X}_i - \bar{X}_{..})^2 = \sum T^2 - TC = \sum_i \frac{X_i^2}{n_i} - \frac{X_{..}^2}{N}$$

$$SC_{Error} \quad \sum_{ij} (X_{ij} - \bar{X}_i)^2 = \sum X^2 - \sum T^2 = \sum_{ij} X_{ij}^2 - \sum_i \frac{X_i^2}{n_i}$$

$$SC_{total} \quad \sum_{ij} (X_{ij} - \bar{X}_{..})^2 = \sum X^2 - TC = \sum_{ij} X_{ij}^2 - \frac{X_{..}^2}{N}$$

Siendo  $N = \sum_i n_i = nt$

En la composición de estas expresiones se ve que hay una lógica, cada sumando elevado al cuadrado está dividido por el número de elementos que han formado esta suma, por ejemplo, si han intervenido  $n_i$  números se dividirá por  $n_i$ ; si la suma fuera de  $N$  individuos, habría que dividir por  $N$ , etc. Es importante comprender esta lógica, porque en la práctica es la que nos va ayudar a realizar cualquier análisis

de varianza, puesto que la descomposición algebraica no se efectúa normalmente, suponiéndose conocida, aplicando directamente las fórmulas anteriores.

### Descomposición de los grados de libertad en un Modelo de una Vía.-

Los grados de libertad se rigen bajo el mismo esquema que las sumas de cuadrados de la fórmula anterior.

En la componente debida a tratamientos se observa que el primer término es un sumatorio que está compuesto por  $t$  sumas y el segundo es una sola expresión, entonces sus  $gl$  son  $t-1$ .

Respecto al término del error y siguiendo el mismo criterio, el primer sumatorio está compuesto por  $N$  sumas y el segundo tiene  $t$  sumandos, por lo que sus  $gl=N-t$ . De una forma análoga se realizan los  $gl$  del total.

Expresándolo todo en un cuadro se tiene

<i>FV</i>	<i>gl</i>	<i>SC</i>
<i>Tratamientos</i>	$t-1$	$\sum_i \frac{X_{i.}^2}{n_i} - \frac{X_{..}^2}{N}$
<i>Error</i>	$N-t$	$\sum_{ij} X_{ij}^2 - \sum_i \frac{X_{i.}^2}{n_i}$
<i>Total</i>	$N-1$	$\sum_{ij} X_{ij}^2 - \frac{X_{..}^2}{N}$

### Parámetros estimados en el Modelo de una vía.-

Para ver qué parámetros son estimados por estas expresiones, se pueden calcular estas sumas a partir del modelo lineal

$$X_{ij} = \mu + T_i + e_{ij}$$

sumando para  $j$

$$X_{i.} = n_i \mu + n_i T_i + e_{i.}; \quad \bar{X}_{i.} = \mu + T_i + \bar{e}_{i.}$$

y sumando para  $i$

$$X_{..} = N \mu + \sum_i n_i T_i + e_{..}; \quad \bar{X}_{..} = \mu + \bar{e}_{..}$$

ya que

$$\sum_i n_i T_i = 0$$

Por lo que, los cuadrados de las desviaciones debidas a los tratamientos es

$$(\bar{X}_i - \bar{X}_{..})^2 = (\mu + T_i + \bar{e}_i - \mu - \bar{e}_{..})^2 = [T_i + (\bar{e}_i - \bar{e}_{..})]^2$$

Y los cuadrados de las desviaciones debidas al error es

$$(\bar{X}_{ij} - \bar{X}_i)^2 = (\mu + T_i + e_{ij} - \mu - T_i - \bar{e}_{..})^2 = (e_{ij} - \bar{e}_{..})^2$$

Sustituyendo estas expresiones en las fórmula definición de la suma de cuadrados, y operando algebraicamente se llega a

$$\begin{aligned} ESC_{Tratamientos} &= (N - t) \sigma_T^2 + (t - 1) \sigma^2 \\ ESC_{Error} &= (N - t) \sigma^2 \end{aligned}$$

Al dividir estas *SC* por los *gl* tendremos los *CM* esperados

$$\begin{aligned} ECM_{Tratamientos} &= n_i \sigma_T^2 + \sigma^2 \\ ECM_{Error} &= \sigma^2 \end{aligned}$$

Si el modelo fuera desequilibrado, sería

$$\begin{aligned} ECM_{Tratamientos} &= \frac{N^2 - \sum_i n_i}{N(t-1)} \sigma_T^2 + \sigma^2 \\ ECM_{Error} &= \sigma^2 \end{aligned}$$

En el caso de que el factor sea de efectos fijos no se puede hablar que varianza de dicho factor, pues los niveles del factor o tratamientos elegidos son los únicos que hay. Si se tiene que repetir la experiencia se volverán a coger los mismos niveles o tratamientos, por lo tanto no hay varianza, lo que hay es sumatorio de los efectos al cuadrado dividido por el número de efectos, es decir

$$\sigma_T^2 = n_i \frac{\sum_i (\bar{X}_i - \bar{X}_{..})^2}{t-1} = n_i \frac{\sum_i T_i^2}{t-1}$$

De todas las maneras, en este manual se seguirá utilizando la notación de la varianza, aunque el factor sea de efectos fijos, por simplicidad de anotación y, por lo tanto, mejor visualización de las estimas de los cuadrados medios; siendo conscientes de que si el factor es de efectos fijos la  $\sigma^2$  se refiere a los sumatorios de los efectos al cuadrado dividido por el número de efectos.

## Prueba de Hipótesis en el modelo de una vía.-

Como ya se ha discutido anteriormente, lo que se quiere probar es si los tratamientos no producen efectos diferenciales o, si por el contrario, los efectos de los tratamientos son significativos.

En el caso de que los efectos de los tratamientos fueran nulos tendríamos que la varianza debida a dichos tratamientos sería cero, es decir,  $\sigma^2_T = 0$ . Y en el caso contrario  $\sigma^2_T > 0$ . Por tanto, las hipótesis a probar son

$$H_0 : \sigma^2_T \leq 0$$

$$H_1 : \sigma^2_T > 0$$

Como se observa, al ser hipótesis de varianzas (cuadrados) las hipótesis siempre serán de la cola derecha, pues nunca habrán valores negativos.

En el caso de que se cumpla la  $H_0$  se tendrá que la estima del cuadrado medio del tratamiento será

$$ECM_{Tratamiento} = n_i \sigma^2_T + \sigma^2 = n_i 0^2 + \sigma^2 = \sigma^2 = ECM_{Error}$$

Es decir, que el cuadrado medio de los tratamientos estimará lo mismo que el cuadrado medio del error.

Y en el caso de que se cumpla la  $H_1$  se tendrá

$$ECM_{Tratamiento} = n_i \sigma^2_T + \sigma^2 > \sigma^2 = ECM_{Error}$$

Es decir, que el cuadrado medio de los tratamientos estimará una cantidad mayor que la del cuadrado medio del error.

Por lo que el cuadrado medio de los tratamientos será mucho mayor que el cuadrado medio del error cuanto mayor sean los efectos de los tratamientos.

Por tanto, conociendo la distribución  $F$ , el estadístico de prueba puede ser

$$F_o = \frac{CM_{Tratamiento}}{CM_{Error}}$$

que se distribuye como  $F_{(t-1, N-t)}$

Esta prueba confirma lo discutido en páginas anteriores, en las que se afirmó que si la variabilidad individual (*error*) es elevada sería más difícil detectar diferencias entre los tratamientos que en el caso en que la variabilidad individual fuera baja (recuérdese lo discutido en el epígrafe *Comparación de Tratamientos*).

Compruébese, así mismo, que el valor esperado de la  $F_o$  para la hipótesis nula ( $\sigma_T^2=0$ ), es la unidad; y para la hipótesis alternativa es un valor superior a uno. El término del numerador, siempre, será igual o superior al del denominador, pues siempre va a ha estimar algún término más. Si alguna vez sale una  $F_o$  inferior a uno, es como consecuencia del error de muestreo, y probablemente se trate de una  $F$  real igual a uno.

Por lo que el análisis completo es

<i>FV</i>	<i>gl</i>	<i>SC</i>	<i>CM</i>	<i>ECM</i>	<i>F<sub>o</sub></i>
<i>Trata</i>	$t-1$	$\sum_i \frac{X_{i.}^2}{n_i} - \frac{X_{..}^2}{N}$	$\frac{SC_{Trata}}{t-1}$	$\sigma^2 + n_i \sigma_T^2$	$\frac{CM_{Trata}}{CM_{Error}}$
<i>Error</i>	$N-t$	$\sum_{ij} X_{ij}^2 - \sum_i \frac{X_{i.}^2}{n_i}$	$\frac{SC_{Error}}{N-t}$	$\sigma^2$	
<i>Total</i>	$N-1$	$\sum_{ij} X_{ij}^2 - \frac{X_{..}^2}{N}$			

Hay que hacer notar que la hipótesis nula,  $\sigma_T^2 = 0$ , equivale a la hipótesis de que las medias de los tratamientos son iguales, es decir

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_t$$

Y la hipótesis alternativa,  $\sigma_T^2 > 0$ , equivale a decir que, al menos, dos tratamientos son diferentes

$$H_1 : \mu_j \neq \mu_h$$

Por lo que la prueba  $F$  del análisis de varianza es una extensión natural de la prueba  $t$  del contraste de dos medias, como ya se ha visto anteriormente, para más de dos medias pero por medio del análisis de la varianza, de ahí su nombre.

### Ejemplo.-

Resolvamos a continuación el ejemplo que se propuso en el epígrafe *Planteamiento del análisis de la varianza*. Se trataba, por ejemplo, de doce vacas sometidas a tres tratamientos cuyas producciones fueron

$T_1$	$T_2$	$T_3$	
5	6	1	
1	2	3	
4	7	0	
2	1	4	
12	16	8	36

$$\sum T^2 = \frac{1}{n} \sum_i X_i^2 = \frac{1}{4} (12^2 + 16^2 + 8^2) = 116$$

$$TC = \frac{X_{..}^2}{N} = \frac{36^2}{3 \times 4} = 108$$

$$\sum X^2 = \sum_{ij} X_{ij}^2 = 5^2 + 1^2 + \dots + 4^2 = 162$$

FV	gl	SC	CM	F <sub>o</sub>	ECM
Trata	3-1	116-108=8	$\frac{8}{2} = 4$	$\frac{4}{5.11} = 0.783ns$	$\sigma^2 + 4\sigma_T^2$
Error	12-3	162-116=46	$\frac{46}{9} = 5.11$		$\sigma^2$
total	12-1	162-108=54			

La F de contraste es  $F_{(2,9; 0.05)} = 4.26$

Al ser la  $F_o$  no significativa, indica que los efectos de los tratamientos no son efectivos. Por lo que se ha concluido lo que se dedujo intuitivamente en el epígrafe *Comparación de Tratamientos*.

#### Archivo del programa SAS (C5-1.SAS).-

```

title 'Análisis de la varianza, una vía';
options ls=75 ps=60;
data vacas;
infile 'c5-1.dat';
input trata $ prod @@;
proc ANOVA;
  class trata;
  model prod=trata;
run;
*Si se quiere saber la estima de los cuadrados medios se puede realizar el
siguiente programa;
proc GLM;
  class trata;
  model prod=trata;
  random trata;
run;

```

#### Archivo de datos (C5-1.DAT).-

T1	5	T2	6	T3	1
T1	1	T2	2	T3	3
T1	4	T2	7	T3	0
T1	2	T2	1	T3	4

**Archivo de resultados (C5-1.LST).-**

Dependent Variable: PROD					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	8.00000000	4.00000000	0.78	0.4860
Error	9	46.00000000	5.11111111		
Corrected Total	11	54.00000000			
	R-Square	C.V.	Root MSE	PROD Mean	
	0.148148	75.35922	2.26078	3.00000	
Source	DF	Type III SS	Mean Square	F Value	Pr > F
TRATA	2	8.00000000	4.00000000	0.78	0.4860
Source	Type III Expected Mean Square				
TRATA	Var(Error) + 4 Var(TRATA)				

En el archivo de salida (C5-1.lst) se observa que el valor de la  $F$  es 0.78 significativa al 0.486, esto es, no significativa, por lo que se concluye que los efectos de los tratamientos no son efectivos.

El coeficiente de determinación  $R^2$  ( $R$ -Square), mide la cantidad de variación de la variable analizada que es debida al modelo planteado. Su rango de es 0 a 1 y viene determinado por la razón entre la  $SC$  debida la modelo y la  $SC$  total. Cuanto mayor sea  $R^2$  mejor es explicada la variabilidad de la variable por el modelo. Como se ve, en este ejemplo, el coeficiente de determinación tiene un valor bajo, 0.148148.

Obsérvese que, en las estimas de las varianzas, da negativa la varianza de los tratamientos,

$$\sigma_T^2 = \frac{4 - 5.1111}{4} = -0.278$$

esto es como consecuencia de lo discutido anteriormente, esto es, que la variabilidad individual es muy elevada comparada con la variabilidad debida a los tratamientos, dando ésta negativa, esto es,  $\sigma_T^2 = 0$ .

**Ejemplo.-**

Recuérdese que en el epígrafe *Comparación de Tratamientos*, se propuso otro ejemplo con los mismos efectos para contrastar con el anterior. Resuélvase también este ejemplo para contrastar que eran ciertas las conclusiones intuitivas.

$T_1$	$T_2$	$T_3$	
3	4	2	
3	3	2	
3	5	1	
3	4	3	
12	16	8	36



$$\sum T^2 = \frac{1}{n} \sum_i X_i^2 = \frac{1}{4} (12^2 + 16^2 + 8^2) = 116$$

$$TC = \frac{X_{..}^2}{N} = \frac{36^2}{3 \times 4} = 108$$

$$\sum X^2 = \sum_{ij} X_{ij}^2 = 3^2 + 3^2 + \dots + 3^2 = 120$$

<i>FV</i>	<i>gl</i>	<i>SC</i>	<i>CM</i>	<i>F<sub>o</sub></i>	<i>ECM</i>
<i>Trata</i>	2	8	4	9.000**	$\sigma^2 + 4 \sigma_T^2$
<i>Error</i>	9	4	0.44		$\sigma^2$
<i>total</i>	11	12			

La *F* de contraste es  $F_{(2,9; 0.05)} = 4.26$

Al ser la  $F_o$  significativa, indica que los efectos de los tratamientos son efectivos; al menos dos de ellos son diferentes. Por lo que se ha concluido lo que se dedujo intuitivamente en el epígrafe *Comparación de Tratamientos*.

Para saber cuales son los tratamientos que son diferentes, véase más adelante el epígrafe *contraste múltiple de medias*.

### Archivo del programa SAS (C5-2.SAS)-

```

title 'Análisis de la varianza, una vía, ejemplo 2';
options ls=75 ps=60;
data vacas;
infile 'c5-2.dat';
input trata $ prod @@;
proc ANOVA;
  class trata;
  model prod=trata;
run;
*Si se quiere saber la estima de los cuadrados medios se puede realizar el
siguiente programa;
proc GLM;
  class trata;
  model prod=trata;
  random trata;
run;

```

### Archivo de datos (C5-2.DAT)-

T1	3	T2	4	T3	2
T1	3	T2	3	T3	2
T1	3	T2	5	T3	1
T1	3	T2	4	T3	3

**Archivo de resultados (C5-2.LST).-**

Dependent Variable: PROD					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	8.00000000	4.00000000	9.00	0.0071
Error	9	4.00000000	0.44444444		
Corrected Total	11	12.00000000			
	R-Square	C.V.	Root MSE		PROD Mean
	0.666667	22.22222	0.66667		3.00000
Source	DF	Type III SS	Mean Square	F Value	Pr > F
TRATA	2	8.00000000	4.00000000	9.00	0.0071
Source	Type III Expected Mean Square				
TRATA	Var(Error) + 4 Var(TRATA)				

En el archivo de salida (c5-2.lst) se observa que el valor de la **F** es 9.00 significativa al 0.0071, esto es, significativa al 0.01 (dos asteriscos), por lo que se concluye que los efectos de los tratamiento son distintos.

También se observa que el coeficiente de determinación (R-Square = 0.66667) es cuatro veces superior que el del ejemplo anterior.

Obsérvese que, en las estimas de las varianzas, la varianza de los tratamientos vale,

$$\sigma_T^2 = \frac{4 - 0.4444}{4} = 0.889$$

y la prueba **F** indica que es significativamente diferente de cero, esto es,  $\sigma_T^2 > 0$ .

**Generalización del ANOVA de una vía a más de una variable.-**

Como ocurría con el ANOVA sin factores, la generalización para más de una variable no tiene ninguna complicación conceptual.

**Ejemplo.-**

Supóngase que el mismo ejemplo anterior en el que a cada unidad se le hubiera medido cinco variables, de manera que la tabla de datos sería

	Tratamiento 1					Tratamiento 2					Tratamiento 3				
	V1	V2	V3	V4	V5	V1	V2	V3	V4	V5	V1	V2	V3	V4	V4
3	10	110	50	82	4	15	120	60	86	1	18	130	71	91	
4	13	112	49	81	3	14	122	62	88	2	16	132	72	90	
2	12	111	51	83	5	16	121	63	84	2	17	131	73	95	
3	9	114	52	80	4	13	119	59	89	3	19	133	74	93	

### Archivo del programa SAS (C5-3.SAS).-

```
title 'Análisis de la varianza, una vía, multivariante';
options ls=75 ps=60;
data vacas;
infile 'C5-3.dat';
input trata $ v1-v5 ;
proc glm;
  class trata;
  model v1-v5 = trata;
  manova h=trata;
run;
```

Si no se desean las salidas univariantes, se pone en el modelo la opción **NOUNI**, (model v1-v5 = trata /nouni;).

### Archivo de datos (C5-3.DAT).-

```
T1 3 10 110 50 82
T1 4 13 112 49 81
T1 2 12 111 51 83
T1 3 9 114 52 80
T2 4 15 120 60 86
T2 3 14 122 62 88
T2 5 16 121 63 84
T2 4 13 119 59 89
T3 1 18 130 71 91
T3 2 16 132 72 90
T3 2 17 131 73 95
T3 3 19 133 74 93
```

### Archivo de resultados (C5-3.LST).-

Se observa que en las salidas univariantes, los tratamientos son significativos para las cinco variables.

Manova Test Criteria and F Approximations for the Hypothesis of no Overall TRATA Effect					
H = Type III SS&CP Matrix for TRATA    E = Error SS&CP Matrix					
S=2    M=1    N=1.5					
Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.00507939	13.0312	10	10	0.0002
Pillai's Trace	1.55554931	4.1999	10	12	0.0109
Hotelling-Lawley Trace	85.50071503	34.2003	10	8	0.0001
Roy's Greatest Root	84.18971010	101.0277	5	6	0.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.  
NOTE: F Statistic for Wilks' Lambda is exact.

En la salida multivariante las pruebas **F** de los cuatro estadísticos son diferentes (el SAS advierte que la **F** de la lambda de *Wiks* es exacta) pero todas ellas

significativas, por lo que se concluye que los efectos de los tratamientos propicia diferentes reacciones en las cinco variables en conjunto.

### **Es suficiente considerar un solo factor.-**

Recordando los ejemplos anteriores se deduce que si se pudieran eliminar los efectos individuales, todas las producciones dentro de un mismo tratamiento serían iguales y las diferencias obtenidas, por pequeñas que fueran, serían debidas a estos tratamientos y, por tanto, serían fácilmente apreciables. Por eso lo importante es controlar al máximo las causas que pueden aumentar dichos efectos individuales; por ejemplo, eligiendo animales genéticamente parecidos en cuanto a productividad y controlando lo mejor posible el medio ambiente de manera que todas las unidades experimentales se encuentren en las mismas condiciones en todo lo que no conciernan a los tratamientos y sobre todo, hay que asegurarse de una aleatorización correcta.

Si existiera cualquier otra causa que aumentara la variabilidad individual hay que tenerla en cuenta como factor o fuente de variación, aunque no interese al problema que se está estudiando, con objeto de que la variabilidad individual se mantenga en niveles que permita detectar, si las hubiera, diferencias entre los tratamientos, por muy pequeñas que estas sean. Si ese otro factor no es considerado en el modelo, puede ocurrir que los tratamientos sean efectivamente diferentes, pero su variabilidad quede oculta en la variabilidad individual, una de cuyas componentes importantes sea el factor no considerado.

Cuando se mide una variable, como puede ser las producciones ganaderas en función de tres dietas; el aumento de peso de niños, también, en función de tres dietas; la cantidad de linfocitos en sangre en función de la aplicación de tres antibióticos, *etc*, ocurrirá, como ya se ha dicho, que están interviniendo otros factores en la variabilidad de dicha variable, muchos de esos factores no son controlables, pero otros factores si lo son. La variabilidad de estos otros factores sin interés para el problema puede interferir en la manifestación estadística de los efectos de los tratamientos, por lo que hay que tenerlos en cuenta. Por ejemplo, puede ocurrir que la experiencia con los tres antibióticos se este realizando en tres plantas del hospital, o en tres hospitales o en tres ganaderías, si se trata de animales domésticos, *etc*. En la experiencia con el crecimiento de niños puede ocurrir que ésta se este haciendo con niños que viven en barrios diferentes (diferentes poderes adquisitivos de las familias), o con niños que tienen diferentes antecedentes clínicos, *etc*.

### **Ejemplo.-**

Por lo expuesto en el párrafo anterior por lo que se va a considerar otro ejemplo en el que se tuvieran los mismos datos del ejemplo primero (página 159), pero en el que se nos hubiera informado que los animales utilizados pertenecían a cuatro líneas distintas, o a cuatro orígenes distintos, o que la experiencia se realizó repartiendo los animales en cuatro corrales, *etc*. Supongamos que son animales de cuatro líneas o razas, y que las producciones obtenidas se podían agrupar así

	$T_1$	$T_2$	$T_3$			
$R_1$	$X_{11}=5$	$X_{21}=7$	$X_{31}=4$	$X_{.1}=16$	$\bar{X}_{.1}=5.3$	$\rho_1=2.3$
$R_2$	$X_{12}=1$	$X_{22}=1$	$X_{32}=0$	$X_{.2}=2$	$\bar{X}_{.2}=0.7$	$\rho_2=-2.3$
$R_3$	$X_{13}=4$	$X_{23}=6$	$X_{33}=3$	$X_{.3}=13$	$\bar{X}_{.3}=4.3$	$\rho_3=1.3$
$R_4$	$X_{14}=2$	$X_{24}=2$	$X_{34}=1$	$X_{.4}=5$	$\bar{X}_{.4}=1.7$	$\rho_4=-1.3$
	$X_{.1}=12$	$X_{.2}=16$	$X_{.3}=8$	$X_{..}=36$		
	$\bar{X}_{.1}=3$	$\bar{X}_{.2}=4$	$\bar{X}_{.3}=2$		$\bar{X}_{..}=3$	
	$\tau_1=0$	$\tau_2=1$	$\tau_3=-1$			

En este caso, el que una unidad experimental pertenezca a una raza o a otra no es aleatorio, sino que es el experimentador el que lo determina con arreglo a criterios ajenos al problema; normalmente esta segunda fuente de variación es debida a factores ambientales o a criterios de clasificación. Pero dentro de cada uno de los niveles de este segundo factor los tratamientos se aplican aleatoriamente. En este ejemplo concreto, se han tenido que elegir aleatoriamente tres animales en cada raza para aplicarle aleatoriamente uno de los tratamientos a cada uno. Por lo tanto, el que un animal pertenezca a una u otra raza no es aleatorio, mientras que los tratamientos han sido aplicados aleatoriamente a cada animal dentro de cada raza.

Recuérdese que cuando se vieron estos mismos datos en un modelo completamente aleatorio, se llegó a la conclusión de que no se podía determinar si había un tratamiento mejor porque la diferencia de los valores dentro de cada tratamiento era muy superior a la diferencia de los valores medios entre tratamientos; sin embargo ahora si se observa que dentro de cada raza siempre es el tratamiento segundo el mejor, por lo que al haber introducido, en el modelo, el factor que cuya variabilidad nos impedía llegar a una conclusión, se visualiza perfectamente las diferencias entre tratamientos. Al ANOVA le ocurre lo mismo.

Tal como se ha hecho con los ejemplos anteriores, la relación existe entre estos valores puede expresarse por medio de desviaciones, de manera que la desviación total de una unidad experimental (valor de dicha unidad menos la media general) es igual a la desviación del grupo en que está dicho individuo (media del grupo menos media general) más la desviación debida a la variación individual (lo que el individuo tiene al final con respecto a lo que tendría por el grupo al que pertenece). Pero en este caso cada individuo pertenece a dos grupos, uno con respecto a  $T_i$  y otro con respecto a  $R_j$ , por lo que la desviación total será igual a la  $i$ -ésima desviación del grupo  $T_i$  más la  $j$ -ésima desviación de grupo  $R_j$  más la desviación individual, es decir

$$(X_{ij} - \bar{X}_{..}) = (\bar{X}_{.i} - \bar{X}_{..}) + (\bar{X}_{.j} - \bar{X}_{..}) + (X_{ij} - \bar{X}_{ij})$$

Al haber una sola medida dentro de cada casilla, el último sumando de esta expresión es cero, pues la media de un número es el mismo número, por lo que no se puede medir la desviación individual o debida al *error*. Pero si no existe relación entre los dos factores, es decir, los efectos de un factor son independientes de los del otro factor (ver más adelante el epígrafe *Interacción*, se puede estimar esta desviación

individual considerando que ésta es igual que la desviación de la unidad experimental en cuestión con respecto a la media general menos las desviaciones de los dos grupos a los que pertenece.

Esto se puede expresar así

$$\begin{aligned} e_{ij} &= (X_{ij} - \bar{X}_{..}) - (\bar{X}_i - \bar{X}_{..}) - (\bar{X}_j - \bar{X}_{..}) = \\ &= X_{ij} - \bar{X}_i - \bar{X}_j + \bar{X}_{..} \end{aligned}$$

es decir, la desviación individual de una unidad experimental es el valor de dicha unidad menos las medias de los dos grupos a los que pertenece mas la media general. Por lo que la desviación total queda de la siguiente manera

$$(X_{ij} - \bar{X}_{..}) = (\bar{X}_i - \bar{X}_{..}) + (\bar{X}_j - \bar{X}_{..}) + (X_{ij} - \bar{X}_i - \bar{X}_j + \bar{X}_{..})$$

Pasando la media general del termino de la derecha al de la izquierda, se obtiene

$$X_{ij} = \bar{X}_{..} + (\bar{X}_i - \bar{X}_{..}) + (\bar{X}_j - \bar{X}_{..}) + (X_{ij} - \bar{X}_i - \bar{X}_j + \bar{X}_{..})$$

Lo que quiere decir que la medida de una unidad experimental es igual a la suma de cuatro sumandos (en el anterior modelo había solo tres sumandos), el primero es la estima de la *media general*,  $\mu$ , el segundo es la estima del efecto del *tratamiento*,  $\tau_i$ , el tercero es la estima del efecto de la *raza*,  $\rho_j$ , y el cuarto es la estima del *error experimental*,  $\varepsilon_{ij}$ . Por lo que en este caso, el modelo lineal ya no es el mismo del ejemplo anterior, ya que hay que añadirle el nuevo factor, el nuevo modelo será

$$\begin{aligned} X_{ij} &= \bar{X}_{..} + (\bar{X}_i - \bar{X}_{..}) + (\bar{X}_j - \bar{X}_{..}) + (X_{ij} - \bar{X}_i - \bar{X}_j + \bar{X}_{..}) \\ X_{ij} &= \mu + T_i + R_j + \varepsilon_{ij} \end{aligned}$$

Ahora, en cada individuo o unidad experimental (dato) influye, además de la media general y el efecto de los *tratamientos*, un efecto *raza* que antes estaba inflando el efecto correspondiente a los efectos incontrolados ( $e$ ). Es decir, si denominados  $e'$  al error de este modelo y  $e$  al error del primer modelo,

$$\begin{aligned} X_{ij} &= \mu + T_i + e_{ij} \\ X_{ij} &= \mu + T_i + R_j + e'_{ij} \\ &\text{se tiene que} \\ e'_{ij} &= e_{ij} - R_j \end{aligned}$$

por lo que el nuevo error será más pequeño cuanto más grande sea la variabilidad debida a la segunda fuente de variación introducida en el modelo y por tanto, se tendrá mayor precisión de medida.

Efectivamente, después de introducir este nuevo factor se observa que, dentro de cada raza, el  $T_2$  es mejor que el  $T_1$  y este es mejor que el  $T_3$  en dos razas, mientras

que en los otros dos  $T_2 = T_1 > T_3$ , por lo que se podría afirmar con cierta seguridad, por lo menos, que el tratamiento de mayor efecto ( $T_2$ ) es superior que el tratamiento de menor efecto ( $T_3$ ). Es decir, para unos mismos datos, si se analizan con una sola fuente de variación no se observan diferencias entre los diferentes niveles de ese factor, mientras que si se analizan teniendo en cuenta otra fuente de variación, si se observan diferencias entre los niveles del factor que antes no se observó diferencias. Esto es como consecuencia de que al introducir el segundo factor se ha aumentado la precisión de medida que ha hecho posible una mayor discriminación de diferencias.

A la segunda fuente de variación, que no tiene interés para el problema planteado, y que se introduce en el modelo con el único objeto de disminuir la variabilidad individual, se le denomina genéricamente como *BLOQUES*. No es la primera vez que se estudia esta cuestión en este manual; anteriormente se estudió en el epígrafe *Comparación de medias muestrales usando observaciones emparejadas* que si las observaciones o medidas se realizan dentro de parejas, se podía tener en cuenta la existencia de esta parejas y realizar las comparaciones dentro de las parejas; eso mismo es lo que se ha hecho en el ejemplo anterior, solo que no eran parejas sino tríos, y las comparaciones se han realizado dentro de los tríos o razas. El análisis estadístico se verá más adelante.

### **Descomposición de la varianza en un modelo factorial sin repetición.-**

La descomposición de una medida cualquiera respecto a la media general, al igual que se hizo anteriormente, puede descomponerse en una serie de sumandos de la siguiente forma

$$X_{ij} - \bar{X}_{..} = (\bar{X}_i - \bar{X}_{..}) + (\bar{X}_j - \bar{X}_{..}) + (X_{ij} - \bar{X}_i - \bar{X}_j + \bar{X}_{..})$$

En el segundo miembro de la igualdad, el primer término mide la desviación del tratamiento  $i$  con respecto a la media general (efecto tratamiento), el segundo término mide la desviación de la raza  $j$  con respecto a la media general (efecto raza) y el tercer término mide la influencia debida a que el elemento  $X_{ij}$  se encuentre en una combinación  $i$ -tratamiento/ $j$ -raza determinada.

Ahora se realizan las mismas operaciones que se hicieron con el modelo de una vía, es decir, elevar al cuadrado esta última expresión y sumar para  $i$  y para  $j$ . Dado que el primer miembro y los dos primeros términos del segundo miembro presentan la misma forma que en el modelo de una vía, no expondremos aquí su cálculo, puesto que los pasos a seguir son completamente análogos. Se anotará directamente el resultado final de la operación en el siguiente cuadro.

$$\begin{aligned}
 SC_{Trata} \quad \sum_{ij} (\bar{X}_{i.} - \bar{X}_{..})^2 &= \frac{1}{r} \sum_i X_{i.}^2 - \frac{X_{..}^2}{N} \\
 SC_{Raza} \quad \sum_{ij} (\bar{X}_{.j} - \bar{X}_{..})^2 &= \frac{1}{t} \sum_j X_{.j}^2 - \frac{X_{..}^2}{N} \\
 SC_{Error} \quad \sum_{ij} (X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..})^2 &= \sum_{ij} X_{ij}^2 - \frac{1}{r} \sum_i X_{i.}^2 - \frac{1}{t} \sum_j X_{.j}^2 + \frac{X_{..}^2}{N} \\
 SC_{total} \quad \sum_{ij} (X_{ij} - \bar{X}_{..})^2 &= \sum_{ij} X_{ij}^2 - \frac{X_{..}^2}{N}
 \end{aligned}$$

Siendo  $N = tr$

En la composición de estas expresiones se ve, como en el modelo de una vía, que hay una lógica, cada sumando elevado al cuadrado está dividido por el número de elementos que han formado esta suma, por ejemplo, si han intervenido  $r$  números se dividirá por  $r$ ; si la suma es de  $t$  individuos, hay que dividir por  $t$ ; y si la suma es de  $N$  individuos, hay que dividir por  $N$ , etc.

### Descomposición de los grados de libertad en un modelo factorial sin repetición.-

Lo mismo que se dijo en el caso del análisis de una sola vía sobre la lógica de la composición de estas expresiones puede aplicarse aquí, así como lo concerniente al concepto de grados de libertad.

Se puede esquematizar en el siguiente cuadro

<i>FV</i>	<i>gl</i>	<i>SC</i>
<i>Tratamientos</i>	$t-1$	$\frac{1}{r} \sum_i X_{i.}^2 - \frac{X_{..}^2}{N}$
<i>Raza</i>	$r-1$	$\frac{1}{t} \sum_j X_{.j}^2 - \frac{X_{..}^2}{N}$
<i>Error</i>	$N-t-r+1$	$\sum_{ij} X_{ij}^2 - \frac{1}{r} \sum_i X_{i.}^2 - \frac{1}{t} \sum_j X_{.j}^2 + \frac{X_{..}^2}{N}$
<i>Total</i>	$N-1$	$\sum_{ij} X_{ij}^2 - \frac{X_{..}^2}{N}$



### Parámetros estimados en un Modelo Factorial sin repetición.-

Si se calculan las sumas de cuadrados a partir del modelo lineal para ver qué estiman estas sumas de cuadrados, al igual que se hizo con el modelo de una vía se obtendría las esperanzas matemáticas de estas sumas de cuadrados que son las siguientes

$$\begin{aligned}ESC_{Tratamientos} &= (t - 1) (r \sigma_T^2 + \sigma^2) \\ESC_{Razas} &= (r - 1) (t \sigma_R^2 + \sigma^2) \\ESC_{Error} &= (r - 1) (t - 1) \sigma^2\end{aligned}$$

Al dividir las estimas de las *SC* por los *gl* se obtienen las de los *CM*

$$\begin{aligned}ECM_{Tratamientos} &= r \sigma_T^2 + \sigma^2 \\ECM_{Razas} &= t \sigma_R^2 + \sigma^2 \\ECM_{Error} &= \sigma^2\end{aligned}$$

### Prueba de hipótesis en un modelo factorial sin repetición.-

Como en el modelo de una vía, las posibles pruebas de hipótesis son las de contrastar si los efectos debidos a los diferentes niveles de los dos factores son o no significativos, por tanto una prueba puede ser

$$\begin{aligned}H_0 : \sigma_T^2 &= 0 \\H_1 : \sigma_T^2 &> 0\end{aligned}$$

y la otra puede ser

$$\begin{aligned}H_0 : \sigma_R^2 &= 0 \\H_1 : \sigma_R^2 &> 0\end{aligned}$$

En el caso de las primeras hipótesis, si  $\sigma_T^2 = 0$

$$ECM_{Tratamiento} = r \sigma_T^2 + \sigma^2 = r \cdot 0^2 + \sigma^2 = \sigma^2 = ECM_{Error}$$

Es decir, que el cuadrado medio de los tratamientos estimará lo mismo que el cuadrado medio del error.

Y si  $\sigma_T^2 > 0$

$$ECM_{Tratamiento} = r \sigma_T^2 + \sigma^2 > \sigma^2 = ECM_{Error}$$

En el caso de las segundas hipótesis, si  $\sigma_R^2 = 0$

$$ECM_{Raza} = t \sigma_R^2 + \sigma^2 = t 0^2 + \sigma^2 = \sigma^2 = ECM_{Error}$$

Es decir, que el cuadrado medio de las razas estimará lo mismo que el cuadrado medio del error.

Y si  $\sigma_R^2 > 0$

$$ECM_{Raza} = t \sigma_R^2 + \sigma^2 > \sigma^2 = ECM_{Error}$$

Por lo que los estadísticos de prueba serán, para la primera prueba

$$F_o = \frac{CM_{Tratamiento}}{CM_{Error}}$$

y para la segunda

$$F_o = \frac{CM_{Raza}}{CM_{Error}}$$

Quedando el análisis de esta manera

FV	gl	SC	CM	ECM	F <sub>o</sub>
Trata	t-1	$\frac{1}{r} \sum_i X_{i.}^2 - \frac{X_{..}^2}{N}$	$\frac{SC_{Trata}}{t-1}$	$\sigma^2 + r\sigma_T^2$	$\frac{CM_{Trata}}{CM_{Error}}$
Raza	r-1	$\frac{1}{t} \sum_j X_{.j}^2 - \frac{X_{..}^2}{N}$	$\frac{SC_{Raza}}{r-1}$	$\sigma^2 + t\sigma_R^2$	$\frac{CM_{Raza}}{CM_{Error}}$
Error	N-t-r+1	$\sum_{ij} X_{ij}^2 - \frac{1}{r} \sum_i X_{i.}^2 - \frac{1}{t} \sum_j X_{.j}^2 + \frac{X_{..}^2}{N}$	$\frac{SC_{Error}}{N-t-r+1}$	$\sigma^2$	
total	N-1	$\sum_{ij} X_{ij}^2 - \frac{X_{..}^2}{N}$			

### Ejemplo.-

Veamos el ejemplo planteado en el epígrafe *Es suficiente considerar un solo factor*. Recuérdese que los datos eran los mismos del ejemplo primero, pero en el que

se nos hubiera informado que los animales utilizados pertenecían a cuatro líneas distintas, o a cuatro orígenes distintos, o que la experiencia se realizó repartiendo los animales en cuatro corrales, etc. Supongamos que son animales de cuatro líneas o razas, y que las producciones obtenidas se podían agrupar así

	$T_1$	$T_2$	$T_3$	
$R_1$	5	7	4	16
$R_2$	1	1	0	2
$R_3$	4	6	3	13
$R_4$	2	2	1	5
	12	16	8	36

$$\frac{1}{r} \sum_i X_i^2 = \frac{1}{4} (12^2 + 16^2 + 8^2) = 116$$

$$\frac{1}{t} \sum_j X_j^2 = \frac{1}{3} (16^2 + 2^2 + 13^2 + 5^2) = 151.3333$$

$$\frac{X_{..}^2}{N} = \frac{X_{..}^2}{N} = \frac{36^2}{3 \times 4} = 108$$

$$\sum_{ij} X_{ij}^2 = 5^2 + 1^2 + \dots + 1^2 = 162$$

FV	gl	SC	CM	$F_o$	ECM
Trata	3-1	116-108	$\frac{8}{2} = 4$	$\frac{4}{0.445} = 8.98^*$	$\sigma^2 + 4\sigma_T^2$
Raza	4-1	151.33-108	$\frac{43.33}{3} = 14.44$	$\frac{14.44}{0.445} = 32.45^{***}$	$\sigma^2 + 3\sigma_R^2$
Error	12-3-4+1	162-116- 151.33+108	$\frac{2.67}{6} = 0.45$		$\sigma^2$
total	12-1	162-108			

Las  $F$  de contraste son, para tratamiento  $F_{(2,6; 0.05)} = 5.14$ , y para raza  $F_{(3,6; 0.05)} = 4.76$

Al ser la  $F_o$  de los tratamientos significativa, indica que los efectos de los tratamientos son efectivos; al menos dos de ellos son diferentes. Por lo que se ha concluido lo mismo que se dedujo intuitivamente en el epígrafe *Es suficiente considerar un solo factor* en el que se estudió el aumento de precisión que se produciría si se introducía un nuevo factor que disminuyera la variabilidad individual. Efectivamente, estos mismos datos analizados sin el factor raza, los tratamientos dan no significativos, mientras que ahora se ve que si lo son. Esto es debido a que antes la variabilidad de raza, incluida en el variabilidad individual, ocultaba la variabilidad de los tratamientos, mientras que ahora al separar el factor raza, que es significativo, de la variabilidad

individual hace que esta sea mucho menor y que este análisis de dos vías detecte las diferencias entre tratamientos que no detectaba el análisis de una vía.

### Archivo del programa SAS (C5-4.SAS).-

```

title 'Análisis de la varianza, dos factores cruzados sin repetición';
options ls=75 ps=60;
data facto;
infile 'C5-4.dat';
input trata $ raza $ prod @@;
proc ANOVA;
  class trata raza;
  model prod=trata raza;
*si se quiere saber las estimas de los cuadrados medios se puede hacer el
siguiente programa;
proc GLM;
  class trata raza;
  model prod=trata raza;
  random   trata raza;
run;

```

### Archivo de datos (C5-4.DAT).-

T1 R1 5	T2 R1 7	T3 R1 4
T1 R2 1	T2 R2 1	T3 R2 0
T1 R3 4	T2 R3 6	T3 R3 3
T1 R4 2	T2 R4 2	T3 R4 1

### Archivo de resultados (C5-4.LST).-

Dependent Variable: PROD						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	5	51.3333333	10.2666667	23.10	0.0007	
Error	6	2.6666667	0.4444444			
Corrected Total	11	54.0000000				
	R-Square	C.V.	Root MSE			PROD Mean
	0.950617	22.22222	0.66667			3.00000
Source	DF	Type III SS	Mean Square	F Value	Pr > F	
TRATA	2	8.0000000	4.0000000	9.00	0.0156	
RAZA	3	43.3333333	14.4444444	32.50	0.0004	
Source	Type III	Expected Mean Square				
TRATA	Var(Error) + 4 Var(TRATA)					
RAZA	Var(Error) + 3 Var(RAZA)					

Al ser la  $F_o$  de los tratamientos significativa ( $F=9.00$ ;  $P>0.016$ ), indica que los efectos de los tratamientos son efectivos; al menos dos de ellos son diferentes. Por lo que se ha concluido lo mismo que se dedujo intuitivamente en el epígrafe *Es suficiente considerar un solo factor* en el que se estudió el aumento de precisión que se produciría si se introducía un nuevo factor que disminuyera la variabilidad individual. Efectivamente, estos mismos datos analizados sin el factor *raza*, los tratamientos son no significativos, mientras que ahora se ve que si lo son. Esto es debido a que, antes, la variabilidad de raza, incluida en el variabilidad individual, ocultaba la variabilidad de

los tratamientos, mientras que ahora al separar el factor raza, que es significativo, de la variabilidad individual hace que esta sea mucho menor y que este análisis de dos vías detecte las diferencias entre tratamientos que no detectaba el análisis de una vía.

Obsérvese que en las componentes de las varianzas, efectivamente, la varianza debida a la raza

$$\sigma_T^2 = \frac{14.4444 - 0.4444}{3} = 4.667$$

es mucho mayor de la debida a los tratamientos

$$\sigma_T^2 = \frac{4 - 0.4444}{4} = 0.889$$

lo que hace que ésta quede oculta en el error si no se separan en el modelo

### Generalización del ANOVA de dos vía a más de una variable.-

Como ocurría con los anteriores ANOVAs, la generalización para más de una variable no tiene ninguna complicación conceptual.

### Ejemplo.-

Supóngase que el mismo ejemplo anterior en el que a cada unidad se le hubiera medido cinco variables, de manera que la tabla de datos sería

	Tratamiento 1					Tratamiento 2					Tratamiento 3				
	V1	V2	V3	V4	V5	V1	V2	V3	V4	V5	V1	V2	V3	V4	V4
R1	5	13	124	52	83	7	15	130	70	96	4	28	140	81	91
R2	1	9	110	39	71	1	10	122	62	88	0	16	132	72	80
R3	4	12	112	51	82	6	16	131	73	94	3	27	141	83	95
R4	2	8	121	30	70	2	11	119	59	89	1	19	133	74	83

### Archivo del programa SAS (C5-5.SAS).-

```

title 'Análisis de la varianza, dos factores cruzados sin repetición,
multivariante';
options ls=75 ps=60;
data fact0;
infile 'c5-5.dat';
input trata $ raza $ v1-v5;
proc glm;
  class trata raza;
  model v1-v5 = trata raza;
  manova h = trata;
  manova h = raza;
run;

```

Si no se desean las salidas univariantes, se pone en el modelo la opción **NOUNI**, (model v1-v5 = trata raza / nouni;).

## Archivo de datos (C5-5.DAT).-

```

T1 R1 5 13 124 52 83
T1 R2 1 9 110 39 71
T1 R3 4 12 112 51 82
T1 R4 2 8 121 30 70
T2 R1 7 15 130 70 96
T2 R2 1 10 122 62 88
T2 R3 6 16 131 73 94
T2 R4 2 11 119 59 89
T3 R1 4 28 140 81 91
T3 R2 0 16 132 72 80
T3 R3 3 27 141 83 95
T3 R4 1 19 133 74 83
    
```

## Archivo de resultados (C5-5.LST).-

Manova Test Criteria and F Approximations for  
the Hypothesis of no Overall TRATA Effect  
H = Type III SS&CP Matrix for TRATA E = Error SS&CP Matrix  
S=2 M=1 N=0

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.00026744	24.0594	10	4	0.0038
Pillai's Trace	1.96508424	33.7684	10	6	0.0002
Hotelling-Lawley Trace	128.55474956	12.8555	10	2	0.0743
Roy's Greatest Root	87.12470358	52.2748	5	3	0.0041

NOTE: F Statistic for Roy's Greatest Root is an upper bound.  
NOTE: F Statistic for Wilks' Lambda is exact.

Manova Test Criteria and F Approximations for  
the Hypothesis of no Overall RAZA Effect  
H = Type III SS&CP Matrix for RAZA E = Error SS&CP Matrix  
S=3 M=0.5 N=0

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.00382395	2.5709	15	5.922518	0.1270
Pillai's Trace	1.60371159	0.9188	15	12	0.5684
Hotelling-Lawley Trace	111.97880428	4.9768	15	2	0.1799
Roy's Greatest Root	110.73171560	88.5854	5	4	0.0003

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

Se observa en los análisis univariantes que los efectos de los dos factores (tratamientos y razas) son significativos en las cinco variables, excepto la variable tres para la que no son significativas las razas.

En el análisis multivariante se observa que los tres tratamientos son diferentes para las cinco variables tomadas en conjunto, mientras que la raza no es significativa.

## Generalización del método para el análisis de modelos factoriales sin repetición con más de dos factores.-

Es importante conocer la mecánica del cálculo y análisis que debe seguirse para diversos tipos de modelos, pues aunque hasta ahora se han mostrado sus expresiones

para el casos de factoriales con dos factores sin repetición, interesa saber la generalización a cualquier número de factores.

En el caso de que se tenga un modelo factorial sin repetición con  $f$  factores, el modelo sería

$$X_{ijk\dots f} = \mu + A_i + B_j + C_k + \dots + F_f + e_{ijk\dots f}$$

Los datos, en el caso de tres factores (unos tratamientos y dos factores ambientales o de clasificación), habría que ordenarlos en tablas de triple entrada de manera que las filas y las columnas de la tabla serían los diferentes niveles de dos factores, y cada casilla contendría el tratamiento y el valor de la variable. En el caso de cuatro factores, el esquema sería el mismo pero cada casilla tendría la anotación del tercer factor y del tratamiento que le corresponde más el valor de la variable.

Por ejemplo, para un factorial triple con tres niveles en cada factor, la tabla de datos podría ser de la siguiente manera

	$B_1$	$B_2$	$B_3$
$A_1$	$T_3$ $X_{211}$	$T_2$ $X_{321}$	$T_1$ $X_{131}$
$A_2$	$T_2$ $X_{112}$	$T_1$ $X_{222}$	$T_3$ $X_{332}$
$A_3$	$T_1$ $X_{313}$	$T_3$ $X_{123}$	$T_2$ $X_{233}$

El modelo lineal que explica estos datos será

$$X_{ijk} = \mu + A_i + B_j + T_k + e_{ijk}$$

El principal obstáculo a la posibilidad de que se nos presente una experiencia como esta es que los efectos ambientales o de clasificación tienen que tener el mismo número de niveles que los tratamientos. Si no es un cuadrado no es correcto realizar esta experiencia porque en ese caso algunos de los tratamientos se quedarían sin medir en alguno de los niveles de los factores, lo que daría una medida incorrecta de los diferentes efectos.

Otro inconveniente para utilizar este modelo es el mismo que se ha visto en el epígrafe anterior, este es, que hay que estar completamente seguro del supuesto de aditividad, es decir, que no existe interacción entre los factores, pues dicha interacción no se puede introducir en el modelo.

Si no se cumplen algunas de estas dos condiciones habría que haber realizado la experiencia con repeticiones, teniendo en este caso un modelo factorial de tres factores con repetición que se estudiará en un próximo epígrafe.

Los datos también se podrían haber ordenado en una lista de  $A \times B \times C$  filas, de manera que las  $f$  primeras columnas hicieran referencia el nivel de cada uno de los  $f$  factores y la última columna serían los datos o el valor de la variable en cada casilla. Para el mismo ejemplo de las tablas anteriores tendríamos algo parecido a esto

$A_1$	$B_1$	$T_1$	$X_{111}$	$A_2$	$B_1$	$T_1$	$X_{211}$
$A_1$	$B_1$	$T_2$	$X_{112}$	$A_2$	$B_1$	$T_2$	$X_{212}$
$A_1$	$B_1$	$T_3$	$X_{113}$	$A_2$	$B_1$	$T_3$	$X_{213}$
$A_1$	$B_2$	$T_1$	$X_{121}$	$A_2$	$B_2$	$T_1$	$X_{221}$
$A_1$	$B_2$	$T_2$	$X_{122}$	$A_2$	$B_2$	$T_2$	$X_{222}$
$A_1$	$B_2$	$T_3$	$X_{123}$	$A_2$	$B_2$	$T_3$	$X_{223}$
$A_1$	$B_3$	$T_1$	$X_{131}$	$A_2$	$B_3$	$T_1$	$X_{231}$
$A_1$	$B_3$	$T_2$	$X_{132}$	$A_2$	$B_3$	$T_2$	$X_{232}$
$A_1$	$B_3$	$T_3$	$X_{133}$	$A_2$	$B_3$	$T_3$	$X_{233}$

Como vemos,  $B_1$  en  $A_1$  es el mismo que en  $A_2$  y el mismo que en  $A_3$ , y es el mismo en los tres tratamientos, etc.

Este es el diseño conocido como *Cuadrado Latino* porque las letras que se ponen en cada casilla, que identifican los tratamientos, son letras latinas. El diseño con cuatro factores sin repetición se denomina *Cuadrado Grecolatino*, como consecuencia de que las letras de cada casilla, que identifican los diferentes tratamientos y las diferentes maneras de repetir los tratamientos, son letras latinas y griegas, respectivamente.

El sistema para calcular los  $gl$  y  $SC$  tiene una lógica común; tan sólo hay que tener en cuenta que cuando se quiera hallar los grados de libertad, los números serán los niveles de los factores y para las sumas de cuadrados, los sumatorios estarán formados por tantos sumandos como niveles tienen los factores.

Los grados de libertad de los factores, al igual que en el modelo de una vía, se obtienen restándole la unidad al número de niveles del factor considerado, y como todos los factores tienen el mismo número de niveles, que son el número de tratamientos, todos tendrán  $t-1$  grados de libertad.

La suma de cuadrados es la suma de cada nivel del factor, elevada al cuadrado, dividida por  $t$  y menos el término de corrección.

Los cuadrados medios estiman, además de su varianza, la del error, por lo que, sea cual sea el tipo de factores, el contraste  $F$  de las tres fuentes de variación siempre se realiza dividiendo por el  $CM$  del error.



Resumamos lo dicho hasta ahora en un cuadro

<i>FV</i>	<i>gl</i>	<i>SC</i>	<i>ECM</i>	<i>F</i>
<i>A</i>	<i>t-1</i>	$\frac{1}{t} \sum_i X_{i..}^2 - TC$	$\sigma^2 + t\sigma_A^2$	$\frac{CM_A}{CM_{Error}}$
<i>B</i>	<i>t-1</i>	$\frac{1}{t} \sum_j X_{.j.}^2 - TC$	$\sigma^2 + t\sigma_B^2$	$\frac{CM_B}{CM_{Error}}$
<i>C</i>	<i>t-1</i>	$\frac{1}{t} \sum_k X_{..k}^2 - TC$	$\sigma^2 + t\sigma_C^2$	$\frac{CM_C}{CM_{Error}}$
<i>Error</i>	$(t-1)(t-2)$	$\sum_{ijk} X_{ijk}^2 - \frac{1}{t} \sum_i X_{i..}^2 - \frac{1}{t} \sum_j X_{.j.}^2 - \frac{1}{t} \sum_k X_{..k}^2 + 2TC$	$\sigma^2$	
<i>total</i>	$t^2-1$	$\sum_{ijk} X_{ijk}^2 - TC$		

Siendo  $TC = \frac{X_{...}^2}{t^2}$

**Ejemplo.-**

Siguiendo con el mismo ejemplo utilizado anteriormente, con el que se ha visto cómo para el análisis de una vía (diseño completamente aleatorio) se suponía que todas las individuos eran iguales ante los tratamientos y cómo se concluía, con el ANOVA, que estos tratamientos son iguales. En el siguiente diseño se vio que las individuos procedían de 4 líneas o estirpes o razas diferentes, se hizo, por lo tanto, 4 bloques y se realizó el análisis factorial sin repetición (diseño de bloques aleatorios), detectando, en este caso, el ANOVA, diferencias entre los tratamientos, como consecuencia de que en el diseño completamente aleatorio la variabilidad de los tratamientos quedaba oculta por la variabilidad de las razas (los bloques) que, como se vio, tiene una variabilidad altamente significativa. Ahora, con los mismos datos, se constata que no sólo se tiene razas diferentes, sino que los individuos están agrupadas en corrales diferentes, por lo que se tiene dos restricciones a la hora de aleatorizar, las razas (filas) y los corrales (columnas); por lo que para impedir que la posible variabilidad entre líneas y la posible variabilidad entre corrales enmascare la variabilidad entre tratamientos, se elabora un cuadrado latino de manera que ningún tratamiento se repita en un mismo corral o en una misma raza.

Como se tienen tres tratamiento tiene que haber tres razas y tres corrales. Para seguir usando el mismo ejemplo que ser viene usando desde el principio del capítulo, tomaremos las tres primeras filas de datos, que serán las tres razas, por lo que los datos en el cuadrado latino quedan

		<i>Corral</i>			
		$C_1$	$C_2$	$C_3$	
<i>Raza</i>	$R_1$	$T_3$ 4	$T_2$ 7	$T_1$ 5	16
	$R_2$	$T_2$ 1	$T_1$ 1	$T_3$ 0	2
	$R_3$	$T_1$ 4	$T_3$ 3	$T_2$ 6	13
		9	11	11	31
$\Sigma T_1=10$		$\Sigma T_2=14$		$\Sigma T_3=7$	

El modelo lineal que explica estos datos será

$$X_{ijk} = \mu + R_i + C_j + T_k + e_{ijk}$$

Los cálculos previos son

$$\Sigma R^2 = \frac{1}{t} \Sigma_i X_{i..}^2 = \frac{1}{3} (16^2 + 2^2 + 13^2) = 143$$

$$\Sigma C^2 = \frac{1}{t} \Sigma_j X_{.j.}^2 = \frac{1}{3} (9^2 + 11^2 + 11^2) = 107.67$$

$$\Sigma T^2 = \frac{1}{t} \Sigma_k X_{...k}^2 = \frac{1}{3} (10^2 + 14^2 + 7^2) = 115$$

$$TC = \frac{X_{...}^2}{t^2} = \frac{31^2}{3^2} = 106.78$$

$$\Sigma X^2 = \Sigma_{ij} X_{ijk}^2 = 4^2 + 7^2 + \dots + 6^2 = 153$$

<i>FV</i>	<i>gl</i>	<i>SC</i>	<i>CM</i>	<i>F</i>
<i>R</i>	3-1	143 - 106.78	$\frac{36.22}{2} = 18.11$	$\frac{18.11}{0.445} = 40.70^*$
<i>C</i>	3-1	107.67 - 106.78	$\frac{0.89}{2} = 0.44$	$\frac{0.44}{0.445} = 1.0ns$
<i>T</i>	3-1	115 - 106.78	$\frac{8.22}{2} = 4.11$	$\frac{4.11}{0.445} = 9.23ns$
<i>Error</i>	(3-1) (3-2)	153-143-107.67-115 + 2x106.78	$\frac{0.89}{2} = 0.445$	
<i>total</i>	$3^2-1$	153 - 106.78		

La  $F$  de contraste, para los tres factores, es  $F_{(2,2; 0.05)}=19$

Se observa que los cuadrados medios de *tratamientos* y de *error* son prácticamente los mismos que en el factorial sin repetición, sin embargo los grados de libertad del error ha disminuido drásticamente, esto hace que este diseño no detecte los efectos de los tratamientos (ver más adelante los *inconvenientes de los cuadrados latinos*).

### Archivo del programa SAS (C5-6.SAS)-

```

title 'Factorial triple sin repeticion';
options ls=75 ps=60;
data clat;
infile 'c5-6.dat';
  do raza=1 to 3;
    do corral=1 to 3;
      input trata $ prod @@; output;
    end;
  end;
proc anova;
  class raza corral trata;
  model prod = raza corral trata;
run;

```

### Archivo de datos (C5-6.DAT)-

```

T3 4 T2 7 T1 5
T2 1 T1 1 T3 0
T1 4 T3 3 T2 6

```

### Archivo de resultados (C5-6.LST)-

Dependent Variable: PROD						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	6	45.3333333	7.5555556	17.00	0.0566	
Error	2	0.8888889	0.4444444			
Corrected Total	8	46.2222222				
	R-Square	C.V.	Root MSE	PROD Mean		
	0.980769	19.35484	0.66667	3.44444		
Source	DF	Anova SS	Mean Square	F Value	Pr > F	
RAZA	2	36.2222222	18.1111111	40.75	0.0240	
CORRAL	2	0.8888889	0.4444444	1.00	0.5000	
TRATA	2	8.2222222	4.1111111	9.25	0.0976	

Como se observa en el archivo de resultados, solo las razas son significativas, si bien los tratamientos son significativa al 10%.

## Inconvenientes del cuadrado latino.-

La principal desventaja del cuadrado latino es que el número de filas, de columnas y de tratamientos tiene que ser el mismo. Así, si hay muchos tratamientos, el número de unidades experimentales pronto se hace impracticable. Los cuadrados latinos deben ser superior, en tamaño, a 4x4 y rara vez se usan cuadrados latinos mayores de 12x12 siendo lo más usual cuadrados latinos de 5x5 a 8x8. Los cuadrados latinos también tienen, hasta cierto punto, la misma desventaja que los bloques aleatorios (dos factores sin repetición), a saber, que el error experimental es probable que aumente con el tamaño del cuadro.

Los cuadrados pequeños proporcionan pocos grados de libertad para la estimación del error; en un cuadrado de 2x2 los  $gl=0$ , en un cuadrado de 3x3 los  $gl=2$  y en un cuadrado de 4x4 los  $gl=6$ . Esto hace que no se pueda usar, individualmente, el cuadrado de 2x2, mientras que se podrían utilizar los cuadrados de 3x3 y 4x4 si se produjera una reducción tan importante en el error, comparado con el diseño de bloques al azar o con el diseño completamente aleatorio como para contrarrestar la pérdida de grados de libertad.

Sin embargo, puede incluirse más de un cuadrado en el mismo experimento. Tres cuadrados de 3x3 (9 repeticiones), proporcionan al error 10  $gl$ , mientras que el error de dos cuadrados de 4x4 tiene 15  $gl$ , si se incluye la interacción *cuadrado/tratamiento* en el error. Pero para los cuadrados de 2x2 se requerirían muchos cuadrados, ya que los grados de libertad del error son uno menos el número de cuadrados.

## Ejecución del diseño.-

El orden del cuadrado latino viene impuesto por el número de tratamientos; si se tiene, por ejemplo, cuatro tratamientos se tendrá que realizar un cuadro de 4x4 y si se tiene tres tratamientos, como los del ejemplo, se tendrá que realizar, como así se ha hecho, un cuadro de 3x3 de la siguiente manera

- 1) Primero se asigna a la primera fila y a la primera columna los tres tratamientos en cualquier orden, aleatorio; por ejemplo,  $T_1, T_2, T_3$ .

$T_1$	$T_2$	$T_3$
$T_2$		
$T_3$		

- 2) La segunda fila se completa con una de las dos permutaciones (elegidas al azar) de las letras restantes, es decir,  $T_2, T_3$ , excluyendo aquellas permutaciones en las que algunas de las letras coincida con la

que tiene en la primera fila. Y de esta manera se rellena todas las filas, teniendo

$T_1$	$T_2$	$T_3$
$T_2$	$T_3$	$T_1$
$T_3$	$T_1$	$T_2$

Los cuadrados como este, que tiene la primera fila y la primera columna en un orden determinado como es  $T_1 T_2 T_3$ , es un cuadrado típico.

3) Una vez obtenido el cuadrado típico se obtiene uno intermedio permutando las  $t-1$  últimas filas

$T_1$	$T_2$	$T_3$
$T_3$	$T_1$	$T_2$
$T_2$	$T_3$	$T_1$

y el cuadrado definitivo se obtiene permutando todas las columnas, con lo que se obtiene

$T_3$	$T_2$	$T_1$
$T_2$	$T_1$	$T_3$
$T_1$	$T_3$	$T_2$

Existen tablas para elaborar los cuadrados latinos, si no se desea realizar el proceso anteriormente descrito. O bien, hoy día se pueden utilizar los paquetes estadísticos para elaborarlos.

### Ejemplo.-

Elabórese el cuadrado latino anterior con el SAS.

**Archivo del programa SAS (C5-7.SAS).-**

```

title 'Elaboración del Diseño de Cuadrados Latinos';
options ls=75 ps=60;
proc plan seed=34859;
  factors raza=3 ordered corral=3 ordered / noprint;
  treatments trata=3 cyclic;
output out=g raza cvals=('raza1' 'raza2' 'raza3') random
      corral cvals=('corral1' 'corral2' 'corral3') random
      trata nvals=(1 2 3) random;
quit;
proc tabulate;
  class raza corral;
  var trata;
  table raza,corral*(trata) / rts=8;
run;

```

**Archivo de resultados (C5-7.LST).-**

Elaboración del Diseño de Cuadrados Latinos

	CORRAL		
	corral1	corral2	corral3
	TRATA	TRATA	TRATA
RAZA			
raza1	3.00	2.00	1.00
raza2	2.00	1.00	3.00
raza3	1.00	3.00	2.00

**Diseños conmutativos o cruzados o sobrecruzados o con intercambio (cross-over).-**

Estos diseños son especialmente adecuados para experimentos con animales grandes (vacas, monos, seres humanos, etc.), con dos o tres tratamientos. En la mayor parte de las investigaciones con seres humanos, ya sea en el campo de la medicina o de la psicología, sólo se consideran unos pocos tratamientos a la vez y en un número de individuos relativamente pequeño, por lo que hay que realizar los diferentes tratamientos en los mismos individuos. En estos casos son especialmente útiles los experimentos conmutativos.

Supóngase que se dispone de doce niños para probar dos tratamientos, por ejemplo  $T_1$  y  $T_2$ . Se podría asignar el tratamiento  $T_1$  a seis de los niños elegidos al azar y el tratamiento  $T_2$  a los otros seis. Esto sería un experimento o *diseño completamente aleatorio*. Sin embargo, normalmente los individuos difieren mucho en su constitución y

en su respuesta a los tratamientos, por lo que se tendría un error experimental grande y, por tanto, el contraste de los tratamientos no es muy preciso. Se puede eliminar la heterogeneidad entre los individuos, utilizando estos como bloques en un *diseño de bloques aleatorios*, siempre que los dos tratamientos sean tales que se puedan aplicar al mismo individuo en tiempos diferentes. Entre los dos tratamientos deberá haber un periodo de reposo para que el efecto del tratamiento que se aplica en primer lugar no interfiera con el efecto del siguiente tratamiento.

Se supondrá que el periodo de reposo entre dos tratamientos es lo suficientemente dilatado como para que los efectos del primer tratamiento desaparezca. Si los efectos de los tratamientos son persistentes (por ejemplo con tratamientos de nutrición) no se podrá utilizar diseños conmutativos. En el siguiente epígrafe se estudiara como se estiman los efectos residuales si los hubiera.

Cuando cada uno de los individuos se emplea como un bloque, el orden de aplicación de los dos tratamientos sobre el individuo deberá ser al azar, con la condición de que haya igual número de individuos que reciban primero el tratamiento  $T_1$  y después el tratamiento  $T_2$ , e individuos que reciben primero el  $T_2$  y después el  $T_1$ . Este diseño recibe el nombre de *diseño conmutativo o con intercambio*.

Supóngase que se tiene dos individuos y dos tratamientos, se necesita, por tanto, dos periodos, en el primero se le aplicara al primer individuo el tratamiento  $T_1$  y al segundo individuo el tratamiento  $T_2$  y en el siguiente periodo lo contrario, es decir

	Individuo 1	Individuo 2
Periodo I	$T_1$	$T_2$
Periodo II	$T_2$	$T_1$

Esto es simplemente un cuadrado latino de  $2 \times 2$ . Por lo que el diseño conmutativo combina características de los bloques aleatorios pequeños con los de los cuadrados latinos pequeños.

En muchas determinaciones biológicas pueden usarse un diseño que se parece mucho al cuadrado latino, pero que puede tener algunas ventajas cuando el número de tratamientos es pequeño. El caso más simple, como se ha visto, es en el que hay dos tratamientos  $T_1$  y  $T_2$ , primero se agrupan las unidades en pares como si se fuera a usar un diseño en bloques al azar. Supóngase que se conoce por experiencias anteriores que una unidad experimental de cada par da una respuesta más alta que la otra, y que la diferencia a favor de la unidad experimental superior sea la misma en todos los pares. Lógicamente, será aconsejable asegurarse de que cada tratamiento se aplique a la mejor unidad experimental en la mitad de las repeticiones y a la peor en la otra mitad. Los pares o repeticiones, que deben ser un número par, se dividen al azar en dos grupos iguales, el primero para recibir el tratamiento  $T_1$  sobre el mejor de cada par y el segundo, para recibir el tratamiento  $T_2$  sobre el mejor.

Pongamos como ejemplo los tiempos necesarios para calcular las sumas de cuadrados de 27 observaciones, en cada una de dos *calculadoras* ( $T_1$  y  $T_2$ ) cuyas velocidades se desean comparar. Se calculan 10 sumas de cuadrados, dos veces

cada una, con sendas calculadoras (10 bloques o 10 niveles del primer factor). Se usa el diseño conmutativo porque se piensa que el segundo cálculo de una suma de cuadrados puede ser más rápido que el primero, de tal manera que para hacer una comparación justa, cada calculadora debe emplearse en primer lugar en sendos grupos de cinco repeticiones. De la aleatorización puede resultar que la calculadora  $T_1$  se use la primera en el grupo de las repeticiones 1, 3, 6, 8 y 9, por ejemplo; y la calculadora  $T_2$  se use en primer lugar en el grupo de las repeticiones restantes. El modelo lineal que describiría más fielmente estos datos sería

$$X_{ijk} = \mu + B_i + G_j + T_k + e_{ijk}$$

Los grados de libertad del análisis de varianza serían

<i>FV</i>	<i>gl</i>
<i>Bloque</i>	9
<i>Grupo</i>	1
<i>Calculadora</i>	1
<i>error</i>	8
<i>Total</i>	19

Como se ve, los grupos sólo tienen 1 *gl*. O sea, que el diseño propuesto elimina del error únicamente la diferencia promedio de todos los pares (bloques). Si la diferencia entre los pares no fuera constante, sino que varía de un bloque a otro, esta variación incrementa el error experimental.

En algunos experimentos se sabe o prevé que la diferencia entre los pares varía de un bloque a otro. Supóngase que se espera que la diferencia disminuya progresivamente de la repetición 1 a la 10. Si esto es cierto, se tiene un diseño más correcto si se usan cinco cuadrados latinos 2x2, como el visto en la página anterior, de la siguiente manera

<i>Cuadrado</i>	I		II		III		IV		V	
<i>Bloque</i>	1	2	1	2	1	2	1	2	1	2
<i>Periodo 1</i>	$T_1$	$T_2$	$T_1$	$T_2$	$T_2$	$T_1$	$T_1$	$T_2$	$T_1$	$T_2$
<i>Periodo 2</i>	$T_2$	$T_1$	$T_2$	$T_1$	$T_1$	$T_2$	$T_2$	$T_1$	$T_2$	$T_1$

De esta manera, se le quita al error los *gl* del factor *Cuadrado*. Eliminandose, no solamente la diferencia promedio, sino también la varianza en la diferencia de cuadrado a cuadrado. El análisis de varianza combinado de los cinco cuadrados sería



<i>FV</i>	<i>gl</i>
<i>Cuadrado</i>	4
<i>Bloque dentro Cuadrado</i>	5
<i>Periodo dentro Cuadrado</i>	5
<i>Calculadora</i>	1
<i>error</i>	4
<i>Total</i>	19

Los 9 *gl* que suman *cuadrados* (4) y *bloques dentro de cuadrados* (5) son exactamente los 9 *gl* de *bloque* en el diseño conmutativo. Debido a la eliminación más completa de los efectos del lugar o *periodo* en que se usa cada calculadora, hay solamente 4 *gl* para el error, en lugar de los 8 *gl* del diseño conmutativo.

El diseño conmutativo es particularmente apropiado cuando la diferencia entre los periodos es de la misma importancia en todas las repeticiones, ya que en este caso el total de la diferencia real entre los periodos se concentra en el único grado de libertad y la varianza del error no es mayor que la de los cuadrados latinos. Pero aún cuando supiéramos que la diferencia entre los periodos es variable, el diseño conmutativo se prefiere en experimentos pequeños en los que se dispone de pocos grados de libertad para el error.

En la producción lechera, por ejemplo, los diseños conmutativos pueden usarse para comparar los efectos de dos raciones alimenticias sobre la cantidad y calidad de leche producida por los individuos. Como la producción de leche varía a lo largo de la lactación, cada ración se prueba en cada uno de los individuos, alimentándolos ya sea durante la primera o la segunda mitad de la lactación, para que cada individuo sea una repetición separada. La producción de leche es superior en la primera mitad de la lactación que en la segunda mitad, por lo que la primera mitad es siempre la mejor.

El diseño conmutativo puede usarse con cualquier número de tratamientos, sujeto a la restricción de que el número de repeticiones debe ser un múltiplo del de tratamientos. Por ejemplo, con tres tratamientos, puede hacerse un diseño a partir de tres cuadrados latinos 3x3, *ABC, BCA, CAB*, donde el orden de las letras significa el periodo donde se aplica cada tratamiento. Cada periodo se aplica al azar a un tercio de las repeticiones.

### **Ejemplo.-**

Se quiere probar dos tratamientos y para ello se dispone solo de doce individuos, por lo que se decide utilizar los mismos individuos para aplicarle ambos tratamientos, dejando pasar un periodo suficientemente grande como para que desaparezca los efectos del primer tratamiento aplicado antes de aplicar el segundo. Como de todas maneras puede producirse un efecto de *aprendizaje* o de *fatiga* se decide realizar un diseño conmutativo de la siguiente manera y con los siguientes datos

Bloque	1	2	3	4	5	6	7	8	9	10	11	12
Periodo 1	$T_2$ 23	$T_2$ 10	$T_1$ 33	$T_2$ 14	$T_1$ 24	$T_1$ 28	$T_1$ 31	$T_1$ 8	$T_2$ 8	$T_2$ 17	$T_2$ 26	$T_1$ 18
Periodo 2	$T_1$ 21	$T_1$ 11	$T_2$ 28	$T_1$ 27	$T_2$ 20	$T_2$ 12	$T_2$ 20	$T_2$ 13	$T_1$ 11	$T_1$ 14	$T_1$ 26	$T_2$ 13

Si se prevé que la diferencia entre los pares varían de un bloque a otro, este mismo experimento se podría haber diseñado como de seis cuadrados latinos 2x2. Quedando el diseño y los datos de la siguiente manera

Cuadrado	1		2		3		4		5		6	
Bloque	1	2	1	2	1	2	1	2	1	2	1	2
Periodo 1	$T_2$ 23	$T_1$ 11	$T_1$ 33	$T_2$ 14	$T_1$ 24	$T_2$ 12	$T_2$ 20	$T_1$ 8	$T_2$ 8	$T_1$ 14	$T_1$ 26	$T_2$ 13
Periodo 2	$T_1$ 21	$T_2$ 10	$T_2$ 28	$T_1$ 27	$T_2$ 20	$T_1$ 28	$T_1$ 31	$T_2$ 13	$T_1$ 11	$T_2$ 17	$T_2$ 26	$T_1$ 18

Según sea un diseño u otro se tendrá un modelo lineal u otro

**Archivo del programa SAS (C5-8.SAS).-**

```

title 'Diseño Conmutativo';
data di_commu;
do periodo = 1 to 2;
do bloque = 1 to 12;
input trata $ val @@; output;
end;
end;
cards;
b 23 b 10 a 33 b 14 a 24 a 28 a 31 a 8 b 8 b 17 b 26 a 18
a 21 a 11 b 28 a 27 b 20 b 12 b 20 b 13 a 11 a 14 a 26 b 13
;
proc anova;
class periodo bloque trata;
model val = periodo bloque trata;
run;
title 'Diseño seis Cuadrados Latinos';
data seis_cl;
do periodo = 1 to 2;
do cuadrado = 1 to 6;
do bloque = 1 to 2;
input trata $ val @@; output;
end;
end;
end;
cards;
b 23 a 11 a 33 b 14 a 24 b 12 b 20 a 8 b 8 a 14 a 26 b 13
a 21 b 10 b 28 a 27 b 20 a 28 a 31 b 13 a 11 b 17 b 26 a 18
;
proc anova;
class cuadrado periodo bloque trata;
model val = cuadrado periodo (cuadrado) bloque (cuadrado) trata;
run;

```

Archivo de resultados (C5-8.LST).-

Dependent Variable: VAL						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	13	1128.00000	86.76923	4.13	0.0153	
Error	10	210.00000	21.00000			
Corrected Total	23	1338.00000				
	R-Square	C.V.	Root MSE	VAL Mean		
	0.843049	24.11882	4.58258	19.0000		
Source	DF	Anova SS	Mean Square	F Value	Pr > F	
PERIODO	1	24.00000	24.00000	1.14	0.3102	
BLOQUE	11	1008.00000	91.63636	4.36	0.0138	
TRATA	1	96.00000	96.00000	4.57	0.0582	
Dependent Variable: VAL						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	18	1237.50000	68.75000	3.42	0.0891	
Error	5	100.50000	20.10000			
Corrected Total	23	1338.00000				
	R-Square	C.V.	Root MSE	VAL Mean		
	0.924888	23.59633	4.48330	19.0000		
Source	DF	Anova SS	Mean Square	F Value	Pr > F	
CUADRADO	5	400.500000	80.100000	3.99	0.0777	
PERIODO (CUADRADO)	6	133.500000	22.250000	1.11	0.4654	
BLOQUE (CUADRADO)	6	607.500000	101.250000	5.04	0.0483	
TRATA	1	96.000000	96.000000	4.78	0.0806	

## Bibliografía

- Afifi, A.A. y Clark, V.* 1984. COMPUTER-AIDED MULIVARIATE ANALYSIS. Ed: Lifetime Learning Publications. Belmont (USA).
- Bisquerra Alzina, R.* 1989. INTRODUCCIÓN CONCEPTUAL AL ANÁLISIS MULTIVARIABLE. Ed: PPU. Barcelona (España).
- Cochran, W.G. y Cox, G.M.* 1965. DISEÑOS EXPERIMENTALES. Ed. TRILLAS. México.
- Cuadras, C.M.* 1981. MÉTODOS DE ANÁLISIS MULTIVARIANTE. Ed:EUNIBAR. Barcelona (España).
- Dagnelie, P.* 1982. ANALYSE STATISTIQUE À PLUSIEURS VARIABLES. Ed: Les Presses Agronomiques De Gembloux. Gembloux (Belgique).
- Gallego, A., García Dorado, A., Carbonell, M., Toro, M.A.* 1982. PROBLEMAS DE BIOESTADÍSTICA Ed. Universidad Complutense. Madrid.
- González López-Valcárcel, B.* 1991. ANÁLISIS MULTIVARIANTE: APLICACIÓN AL ÁMBITO SANITARIO. Ed: SG Editores. Barcelona (España).
- Judez Asensio, L.* 1989. TÉCNICAS DE ANÁLISIS DE DATOS MULTIDIMENSIONALES. Ed: MAPA. Madrid (España).
- Lebart, L., Morineau, A. Y Fénelon, J.P.* 1979. TRAITEMENT DES DONNÉES STATISTIQUES. Ed: Dunod. Paris (France).
- Lefebvre, J.* 1980. INTRODUCTION AUX ANALYSES STATISTIQUES MULTIDIMENSIONNELLES. Ed: Masson. Paris (France)
- Li, Ching Chun.* 1969. INTRODUCCIÓN A LA ESTADÍSTICA EXPERIMENTAL. Ed. Omega. Barcelona.
- Lite, TM, y Jackson Hills, F.* 1987. MÉTODOS ESTADÍSTICOS PARA LA INVESTIGACIÓN EN LA AGRICULTURA. Ed TRILLAS. México.
- Littel, R.C., Freund, R.J. and Spector, P.C.* 1991. SAS® FOR LINEAR MODELS. SAS Institute Inc., Cary, NC, USA.
- Mills, F.S.* 1969. MÉTODOS ESTADÍSTICOS. Ed. Aguilar. Madrid.
- Milton, J.S.* 1994. ESTADÍSTICA PARA BIOLOGÍA Y CIENCIAS DE LA SALUD. Ed. Interamericana-McGraw-Hill. México.
- Orozco, F.* APUNTES DE ESTADÍSTICA. I.N.I.A. Madrid.
- Ostle, B.* 1965. ESTADÍSTICA APLICADA. Ed. Limusa-Wiley. México.
- Ruiz Maya, L.* 1983. MÉTODOS ESTADÍSTICOS DE INVESTIGACIÓN. Ed. INE. Madrid.
- Snedecor, G.W. y Cochran, W.G.* 1971. MÉTODOS ESTADÍSTICOS. Ed. C.E.C.S.A. México.
- Sokal, R.R. y Rohlf, F.J.* 1994. BIOMETRY. Ed. W.H.FREEMAN. San Francisco.
- Spiegel M.R.* 1990. ESTADÍSTICA. Ed. McGraw-Hill. Madrid.
- Srivastava, M.S. y Carter, E.M.* 1983. AN INTRODUCTION TO APPLIED MULTIVARIATE STATISTICS. Ed:Elsevier Science Publishing. New York (USA).
- Steel, R.* 1996. PRINCIPLES AND PROCEDURES OF STATISTICS. Ed. McGRAW-HILL Education. New York .
- SAS Institute Inc. 1990. SAS/STAT USER'S GUIDE. Volume 1 and 2. Cary, NC, USA.



## **CAPÍTULO 6**

# **Supuestos paramétricos y Transformaciones**



# Supuestos paramétricos y Transformaciones

### Supuestos del análisis de varianza.-

Los supuestos sobre los que se basa un análisis de varianza son, en resumen

- 1) Los tratamientos y los efectos ambientales son aditivos.
- 2) Los efectos del error son aleatorios, independientes y con distribución normal.
- 3) Las varianzas de las diferentes submuestras son homogéneas.
- 4) Las varianzas y las medias de las distintas submuestras no están correlacionadas.

Pero el supuesto fundamental en todo análisis de varianza es que el muestreo de las unidades experimentales y la asignación a los diferentes tratamientos es aleatorio. La no aleatorización o selección muestral puede muy bien reflejarse en falta de independencia de los datos o en heterogeneidad de las varianzas o en la no normalidad de la distribución.

Pasemos a estudiar cada uno de estos supuestos.

### Aleatorización.-

La función de la aleatorización consiste en asegurarse que obtengamos una estima válida e insesgada del error experimental, de las medias de los tratamientos y de las diferencias entre las mismas.



Como es de esperar, el tipo de inferencia estadística que puede hacerse de un grupo de datos depende de la naturaleza de los mismos. Es fácil hacer un experimento de manera que no puedan sacarse inferencias útiles.

La aleatorización es la característica más moderna de los diseños experimentales. Se pueden encontrar experimentos realizados en siglos pasados que incluyen los principios considerados hoy como buenos, con excepción de la aleatorización.

Las ocasiones en que se requiere la aleatorización varían con el tipo del experimento y esta sujeto al juicio del experimentador. Ya hemos visto el experimento en el que se quiere investigar los efectos de tres dietas en el peso. Ya que, evidentemente, no todos los individuos crecen con el mismo ritmo, se puede ver favorecido un tratamiento que resulte asignado a un grupo de individuos de crecimiento rápido. En consecuencia, las dietas se distribuyen en cualquier repetición al azar a los individuos que van a recibirlas, con una nueva distribución al azar en cada repetición. Esta es una *aleatorización general*.

De manera similar, si se están comparando estos tres tratamientos en cuatro razas, éstos deben asignarse al azar en cada una de las razas para cualquier repetición. Esta es una *aleatorización restringida* a las razas.

Algunas veces ésta es la única aleatorización requerida, pero frecuentemente se presentan otras operaciones en el curso del experimento que son también fuentes potenciales de sesgo. Se ha visto que en una operación iterativa, el orden de los eventos puede ser importante, bien sea porque se trate de un proceso de aprendizaje, el cual tienda a hacer las operaciones posteriores mejores que las primeras, o bien a causa de que la fatiga actúa en dirección opuesta. En este caso los sesgos sistemáticos se han evitado mediante la aleatorización del orden de los diferentes tratamientos dentro de cada repetición. En otro caso puede introducir variación los materiales utilizados. Tal es el ejemplo en el que se disponía de tres razas y tres corrales para probar los tres tratamientos, en este caso no siempre se usara el mismo corral para el mismo tratamiento con objeto de evitar la introducción de sesgos a causa de diferencias sistemáticas entre los corrales. En su lugar, los tratamientos se han asignado al azar a cada raza y cada corral. Así, se tiene dos aleatorizaciones, una para asignar cada tratamiento a las razas y otra para asignar la raza a los corrales.

La aleatorización es en cierta forma análoga a un *seguro*, por el hecho de que es una precaución contra interferencias que pueden o no ocurrir, y ser o no importantes si ocurren. Generalmente, es aconsejable tomarse el trabajo de aleatorizar, aun cuando no se espere que haya un sesgo importante al dejar de hacerlo. De esta manera el experimento queda protegido contra sucesos que puedan alterar la esperanza de la estima.

Hay que añadir algo más. Ya se realice el experimento propuesto por medio de una *aleatorización general* o por medio de una *aleatorización restringida*, ambos métodos nos proporcionarán un contraste válido para la prueba de significación. Sin embargo, los análisis que se hagan para esta prueba de significación serán diferentes en los dos casos, pues las fuentes de variación son diferentes y por lo tanto son

diferentes los modelos lineales de ambas experiencias, tal como se ha visto en su momento.

Este es sólo un ejemplo que nos muestra cómo el modo de realizar un experimento determina no sólo si se pueden hacer inferencias, sino también el modelo lineal requerido para realizar dichas inferencias. Por lo tanto hay que asegurarse que el modelo lineal utilizado sea el apropiado para el experimento.

### Aditividad de los tratamientos y de los efectos ambientales.-

Para entender esta cuestión véase en el siguiente capítulo el epígrafe relativo a la *Interacción*.

No se tendrá aditividad cuando los efectos dentro de cada casilla, es decir, los efectos de *tratamiento/bloque* o los de la combinación de los niveles de dos factores, son multiplicativos. Esta falta de aditividad se traduce en lo que en estadística se conoce como *interacción* (verlo más adelante), es decir, que los efectos de los diferentes niveles de un factor no son los mismos en todos los niveles del otro factor, esto es, existen fenómenos de *sinergia* o de *interferencia* entre factores o entre tratamientos bloques.

Supóngase que se tiene dos tratamientos y dos bloques o ambientes, se tendrá un modelo aditivo o multiplicativo cuando se tenga datos de la manera expresada en la siguiente tabla, omitiendo los errores experimentales.

	<i>Aditivo</i>		<i>Multiplicativo</i>		<i>Logaritmos</i>	
	<i>B1</i>	<i>B2</i>	<i>B1</i>	<i>B2</i>	<i>B1</i>	<i>B2</i>
<i>Tratamiento 1</i>	10	20	10	20	1.00	1.30
<i>Tratamiento 2</i>	30	40	30	60	1.48	1.78

Para el modelo aditivo, el aumento del bloque 1 al bloque 2 es una cantidad fija independientemente del tratamiento; lo mismo ocurre para los tratamientos. Cuando los efectos son multiplicativos los aumentos no ocurren de esta manera, aunque los logaritmos de los datos si manifiestan los efectos de manera aditiva, por lo que sería una buena transformación (ver más adelante el epígrafe *Transformación logarítmica*).

La presencia de no aditividad en los datos conduce a una aparente heterogeneidad del error debido a supuestos falsos. Las componente de la varianza del error aportadas por las diferentes observaciones no dan estimaciones de una varianza común. La varianza del error puede ser un tanto ineficaz para hacer comparaciones específicas de medias entre tratamientos aunque la prueba *F* en la que entran todas las medias de los tratamientos se ve muy poco afectada. Pero hay que tener en cuenta que en el caso de comprobarse esa falta de aditividad (como se verá en el siguiente epígrafe con la *prueba de no aditividad* de *Tukey*) los resultados de la prueba *F* puede no tener mucho sentido globalmente, y tal vez sea más apropiado realizar un estudio más pormenorizado por casillas.

**Prueba de no aditividad.-**

En el caso de tener un modelo factorial con repetición o de bloques aleatorios con interacción, esta falta de aditividad se puede resolver introduciendo dicha interacción en el modelo como si fuera un factor más, conservando, por tanto, de esta manera, la aditividad. Pero si se tiene un modelo factorial sin repetición o de bloques aleatorios con una sola medida por casilla, no se puede contrastar la existencia de la interacción pues es la SC de la interacción la que se utiliza como error, por lo tanto no se puede introducir la interacción en el modelo ni se puede probar la existencia de la dicha interacción. Para resolver esta dificultad se ha ideado una *prueba de no aditividad* (de Tukey) consistente en descomponer la suma de cuadrados del error en dos componentes: la suma de cuadrados debida a la aditividad (N) y la suma de cuadrados residual (R).

Supóngase que se tiene el mismo experimento factorial sin repetición visto anteriormente

		Tratamientos			
		T <sub>1</sub>	T <sub>2</sub>	T <sub>3</sub>	Σ
Bloques	R <sub>1</sub>	X <sub>11</sub>	X <sub>21</sub>	X <sub>31</sub>	X <sub>·1</sub>
	R <sub>2</sub>	X <sub>12</sub>	X <sub>22</sub>	X <sub>32</sub>	X <sub>·2</sub>
	R <sub>3</sub>	X <sub>13</sub>	X <sub>23</sub>	X <sub>33</sub>	X <sub>·3</sub>
	R <sub>4</sub>	X <sub>14</sub>	X <sub>24</sub>	X <sub>34</sub>	X <sub>·4</sub>
	Σ	X <sub>1.</sub>	X <sub>2.</sub>	X <sub>3.</sub>	X <sub>..</sub>

Se ha visto que el análisis de la varianza es

FV	gl	SC	CM	ECM	F <sub>o</sub>
Trata	t-1	$\frac{1}{r} \sum_i X_{i.}^2 - \frac{X_{..}^2}{N}$	$\frac{SC_{Trata}}{t-1}$	$\sigma^2 + r\sigma_T^2$	$\frac{CM_{Trata}}{CM_{Error}}$
Raza	r-1	$\frac{1}{t} \sum_j X_{.j}^2 - \frac{X_{..}^2}{N}$	$\frac{SC_{Raza}}{r-1}$	$\sigma^2 + t\sigma_R^2$	$\frac{CM_{Raza}}{CM_{Error}}$
Error	N-t-r+1	$\sum_{ij} X_{ij}^2 - \frac{1}{r} \sum_i X_{i.}^2 - \frac{1}{t} \sum_j X_{.j}^2 + \frac{X_{..}^2}{N}$	$\frac{SC_{Error}}{N-t-r+1}$	$\sigma^2$	
total	N-1	$\sum_{ij} X_{ij}^2 - \frac{X_{..}^2}{N}$			

En el que no se puede probar la interacción, por lo que en caso de sospechar la falta de aditividad se realiza la siguiente prueba ideada por Tukey

La suma de cuadrados debido a la aditividad es

$$SC_{(N)} = \frac{\left[ \sum_{ij} (X_{ij} X_i X_j) - X_{..} \left( \frac{X_{..}^2}{N} + SC_T + SC_R \right) \right]^2}{tr SC_T SC_R}$$

La suma de cuadrados residual es

$$SC_{(R)} = SC_{(Error)} - SC_{(N)}$$

El contraste se hace con la prueba F

$$F = \frac{SC_{(N)}}{\frac{SC_{(R)}}{N-t-r}}$$

que se distribuye como  $F_{(1, (t-1)(r-1))}$ .

Pudiéndose, por tanto, realizar un análisis más completo

FV	gl	SC	CM	F
Trata	t-1	$\frac{1}{r} \sum_i X_i^2 - \frac{X_{..}^2}{N}$	$\frac{SC_{Trata}}{t-1}$	$\frac{CM_{Trata}}{CM_{Error}}$
Raza	r-1	$\frac{1}{t} \sum_j X_j^2 - \frac{X_{..}^2}{N}$	$\frac{SC_{Raza}}{r-1}$	$\frac{CM_{Raza}}{CM_{Error}}$
Aditiv	N-t-r	$SC_{(N)} = \frac{\left[ \sum_{ij} (X_{ij} X_i X_j) - X_{..} \left( \frac{X_{..}^2}{N} + SC_T + SC_R \right) \right]^2}{tr SC_T SC_R}$	$CM_N$	$\frac{CM_N}{CM_{Resid}}$
Resid	N-t-r	$SC_{(Error)} - SC_{(N)}$	$\frac{SC_{Residuo}}{N-t-r+1}$	
Error	N-t-r+1	$\sum_{ij} X_{ij}^2 - \frac{1}{r} \sum_i X_i^2 - \frac{1}{t} \sum_j X_j^2 + \frac{X_{..}^2}{N}$	$\frac{SC_{Error}}{N-t-r+1}$	
Total	N-1	$\sum_{ij} X_{ij}^2 - \frac{X_{..}^2}{N}$		

### Ejemplo.-

Veamos el mismo ejemplo. Se tiene tres *tratamientos* y cuatro *bloque* o razas de la siguiente manera

		Tratamientos			
		$T_1$	$T_2$	$T_3$	$X_j$
Razas	$R_1$	5	7	4	16
	$R_2$	1	1	0	2
	$R_3$	4	6	3	13
	$R_4$	2	2	1	5
$X_i$		12	16	8	36

$$\frac{1}{r} \sum_i X_i^2 = \frac{1}{4} (12^2 + 16^2 + 8^2) = 116$$

$$\frac{1}{t} \sum_j X_j^2 = \frac{1}{3} (16^2 + 2^2 + 13^2 + 5^2) = 151.3333$$

$$\frac{X_{..}^2}{N} = \frac{36^2}{3 \times 4} = 108$$

$$\sum_{ij} X_{ij}^2 = 5^2 + 1^2 + \dots + 1^2 = 162$$

$$\sum_{ij} X_{ij} X_i X_j = 5 \times 16 \times 12 + \dots + 1 \times 5 \times 8 = 5824$$

FV	gl	SC	CM	F
Trata	2	8	4	8.89*
Raza	3	43.33	14.44	32.45***
Aditividad	1	1.8668	1.8668	11.62*
Residuo	5	0.8032	0.1606	
Error	6	2.67	0.45	
Total	11	54		

Luego se acepta la hipótesis de no aditividad o interacción. Habría que interpretar los resultados estadísticos con cierta precaución.

### Normalidad del error.-

Afortunadamente, las desviaciones del supuesto de normalidad no afectan muy seriamente la validez del análisis de varianza. Resulta bastante fútil aplicar estas pruebas como análisis previo a unos datos tomados para un ANOVA, a menos que el número de submuestras con las que estamos trabajando sea muy grande. Sí son útiles para estudios, ad hoc y previos, de normalidad de una nueva variable.

Únicamente una distribución muy asimétrica tendría un efecto marcado sobre el nivel de significación de la prueba  $F$  o sobre la eficacia del diseño. Como una aproximación, algunos autores indican que con la no normalidad en los errores la verdadera probabilidad correspondiente al nivel de significación de 0.01 de la tabla  $F$

puede estar situada entre el 0.005 y el 0.02; y la probabilidad verdadera correspondiente al  $\alpha=0.05$  puede estar entre el 0.04 y el 0.07. Puesto que el efecto general de la no normalidad de los errores es conducir a la aceptación de demasiados resultados como significativos, es razonable ser moderado en cuanto a tal aceptación en el caso de que exista duda en lo que se refiere a la normalidad de la población muestreada.

Muchas veces la falta de normalidad se corrige realizando una transformación como veremos más adelante.

### Pruebas de normalidad con una sola muestra.-

Si se tiene una variable cuantitativa se puede agrupar en clases y realizar la prueba de bondad de ajuste a la distribución normal. Estas pruebas de bondad de ajuste pueden aplicarse a cualquier distribución (*Normal, Binomial, Poisson, etc.*). En el Capítulo 20, dedicado a los **Datos Categóricos** se verán ejemplos con las otras distribuciones.

### Ejemplo.-

Se tiene un carácter productivo medido en una ganadería de 200 individuos y se quiere saber si la variable continua, con la que se mide dicho carácter, se ajusta a una distribución normal.

	Clases	Marcas	O	e	$\frac{(o-e)^2}{e}$	$o \ln\left(\frac{o}{e}\right)$	
	<	14.92	13.575	1	0.69	0.1393	0.3711
	14.92 -	17.61	16.265	2	2.02	0.0002	-0.0199
	17.61 -	20.30	18.955	5	5.86	0.1262	-0.7936
	20.30 -	22.99	21.645	14	13.40	0.0269	0.6132
	22.99 -	25.68	24.335	22	24.18	0.1965	-2.0786
	25.68 -	28.38	27.025	39	34.46	0.5981	4.8267
	28.38 -	31.07	29.727	40	38.78	0.0384	1.2390
	31.07 -	33.76	32.415	28	34.46	1.2110	-7.8886
	33.76 -	36.45	35.105	25	24.18	0.0278	0.8337
	36.45 -	39.14	37.795	13	13.40	0.0119	-0.3940
	39.14 -	41.83	40.485	10	5.86	2.9248	5.3443
	41.83 -	44.52	43.175	0	2.02	2.0200	0.0000
	44.52 >	45.865	45.865	1	0.69	0.3100	0.3711
		Total	200	200.00	7.4605	4.5005	

Los valores esperados se han obtenido de la tabla Z (Tabla 1) usando la estima de los parámetros  $\bar{X} = 29.72$  y  $S^2 = 30.04$

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i} = 7.460ns$$

$$G = 2 \sum_i o_i \ln\left(\frac{o}{e}\right) = 2 \times 4.5005 = 9.001ns$$

$$gl = 13(\text{número clases}) - 1 - 1(\text{estima } \mu) - 1(\text{estima } \sigma) = 10$$

$$\chi^2_{(10; 0.05)} = 18.307$$

El ajuste a la normal es bueno.

### Prueba de Kolmogorov-Smirnov para una muestra.-

Una prueba no paramétrica aplicable a distribuciones de frecuencias continuas divididas en clases y que en muchos casos posee mayor potencia que la prueba  $\chi^2$  o G para bondad de ajuste, es la prueba se *Kolmogorov-Smirnov*.

Esta prueba es de fácil de realización. Está basada en las diferencias absolutas de las distribuciones de frecuencias acumuladas observadas con respecto a las esperadas, expresado en frecuencias relativas. La diferencia máxima relativa se contrasta con el valor crítico

$$D_\alpha = \sqrt{\frac{-\ln \frac{\alpha}{2}}{\frac{2}{n}}}$$

### Ejemplo.-

Realícese el mismo ejemplo anterior.

Clases		<i>o</i>	<i>e</i>	<i>d</i> =   <i>o</i> - <i>e</i>	
<	14.92	1	0.69	0.31	
14.92	- 17.61	3	2.71	0.29	
17.61	- 20.30	8	8.57	0.57	
20.30	- 22.99	22	21.97	0.03	
22.99	- 25.68	44	46.15	2.15	
25.68	- 28.38	83	80.61	2.39	
28.38	- 31.07	123	119.39	3.61	← <i>d</i> <sub>max</sub>
31.07	- 33.76	151	153.85	2.85	
33.76	- 36.45	176	178.03	2.03	
36.45	- 39.14	189	191.43	2.43	
39.14	- 41.83	199	197.29	1.71	
41.83	- 44.52	199	199.31	0.31	
44.52	>	200	200.00	0.00	

La diferencia máxima relativa observada es

$$d_{\max} = \frac{3.61}{200} = 0.018ns$$
$$D_{0.05} = \sqrt{\frac{\frac{-\ln(0.05/2)}{2}}{200}} = 0.096$$

Como la diferencia máxima relativa observada es inferior que el valor crítico, no se puede rechazar la hipótesis nula, por lo que se puede concluir que el ajuste a la normal es bueno.

En la práctica no hay que hallar las diferencias absolutas de todas las clases, sino que hallando primero el valor crítico basta encontrar una diferencia mayor que este valor para que sea significativa la prueba.

### Ejemplo SAS.-

Esta prueba la realiza el SAS en el procedimiento **UNIVARIATE** con la opción **NORMAL** (ver siguiente ejemplo SAS), cuando el número de datos es superior a 3000.

### Prueba de *Shapiro Wilk* para varias muestras.-

La prueba *W* de *Shapiro* y *Wilk* es una prueba de ajuste a normalidad muy útil en muestras pequeñas ( $n < 50$ ). Esta prueba consiste en lo siguiente

Sea una muestra de  $n$  elementos, ordenados de menor a mayor. Se calcula en cada muestra la suma de cuadrados (*SC*) y la  $b$  siguiente

$$b = \sum_i^k a_{n-i+1} (X_{n-i+1} - X_i)$$

siendo

$$k = \frac{n}{2} \quad \text{si } n \text{ es par}$$

ó

$$k = \frac{n-1}{2} \quad \text{si } n \text{ es impar}$$

Los coeficientes  $a_{n-i+1}$  se extraen de la tabla 9.

La *W* se obtiene

$$W_o = \frac{b^2}{SC}$$



Este valor de  $W$  debe ser mayor de cero y menor o igual que uno y se contrasta con el valor tabular de la Tabla 10, considerándose significativo, y por lo tanto rechazando la hipótesis de normalidad, cuando es menor que el valor de las tablas.

**Ejemplo.-**

Se tiene un ANOVA de una vía, por lo que se está suponiendo la normalidad de la variable. Compruébese si es cierto.

<i>Tratamientos</i>				
1	2	3	4	5
0.481	0.504	0.496	0.487	0.603
0.502	0.507	0.502	0.492	0.693
0.509	0.670	0.503	0.496	0.700
0.513	0.684	0.511	0.500	0.713
0.518	0.697		0.506	
	0.713		0.515	

<i>Tratamiento 1</i>				
$a_{n+i+1}$	$X_{n+i+1}$	$b$	$SC$	$W$
0.6646	0.037			
0.2413	0.011	0.0272445	0.0008332	0.891 $_{ns}$

<i>Tratamiento 2</i>				
$a_{n+i+1}$	$X_{n+i+1}$	$b$	$SC$	$W$
0.6431	0.209			
0.2806	0.190			
0.0875	0.014	0.1889469	0.0468948	0.761*

<i>Tratamiento 3</i>				
$a_{n+i+1}$	$X_{n+i+1}$	$b$	$SC$	$W$
0.6872	0.015			
0.1677	0.001	0.0104757	0.0001140	0.963 $_{ns}$

<i>Tratamiento 4</i>				
$a_{n+i+1}$	$X_{n+i+1}$	$b$	$SC$	$W$
0.6431	0.028			
0.2806	0.014			
0.0875	0.004	0.0222852	0.0005073	0.979 $_{ns}$

<i>Tratamiento 5</i>				
$a_{n+i+1}$	$X_{n+i+1}$	$b$	$SC$	$W$
0.6872	0.110			
0.1677	0.007	0.0767659	0.0075568	0.780 $_{ns}$

Se acepta, por tanto, la hipótesis de normalidad en cuatro de las cinco submuestras, si bien en la muestra que se rechaza la hipótesis de normalidad, en la segunda, sobrepasa ligeramente el valor crítico.

**Archivo del programa SAS (C6-1.SAS)-**

```

title 'prueba de normalidad';
options ls=75 ps=60;
data uni;
infile 'c6-1.dat';
input muestra $ valor @@;
proc univariate normal;
var valor;
by muestra;
run;

```

**Archivo de datos (C6-1.DAT)-**

1	0.481	1	0.502	1	0.509	1	0.513	1	0.518		
2	0.504	2	0.507	2	0.670	2	0.684	2	0.697	2	0.713
3	0.496	3	0.502	3	0.503	3	0.511				
4	0.487	4	0.492	4	0.496	4	0.500	4	0.506	4	0.515
5	0.603	5	0.693	5	0.700	5	0.713				

**Archivo de resultados (C6-1.LST)-**

prueba de normalidad			
----- MUESTRA=1 -----			
Univariate Procedure			
Variable=VALOR			
Moments			
N	5	Sum Wgts	5
Mean	0.5046	Sum	2.523
Std Dev	0.014433	Variance	0.000208
Skewness	-1.39676	Kurtosis	2.021156
USS	1.273939	CSS	0.000833
CV	2.860206	Std Mean	0.006454
T:Mean=0	78.17855	Pr> T	0.0001
Num ^= 0	5	Num > 0	5
M(Sign)	2.5	Pr>= M	0.0625
Sgn Rank	7.5	Pr>= S	0.0625
W:Normal	0.890858	Pr<W	0.3590
----- MUESTRA=2 -----			
Univariate Procedure			
Variable=VALOR			
Moments			
N	6	Sum Wgts	6
Mean	0.629167	Sum	3.775
Std Dev	0.096845	Variance	0.009379
Skewness	-0.87525	Kurtosis	-1.86866
USS	2.421999	CSS	0.046895
CV	15.39259	Std Mean	0.039537
T:Mean=0	15.91343	Pr> T	0.0001
Num ^= 0	6	Num > 0	6

M(Sign)	3	Pr>= M	0.0313
Sgn Rank	10.5	Pr>= S	0.0313
W:Normal	0.761298	Pr<W	0.0257
----- MUESTRA=3 -----			
Univariate Procedure			
Variable=VALOR			
Moments			
N	4	Sum Wgts	4
Mean	0.503	Sum	2.012
Std Dev	0.006164	Variance	0.000038
Skewness	0.478126	Kurtosis	1.5
USS	1.01215	CSS	0.000114
CV	1.22553	Std Mean	0.003082
T:Mean=0	163.1947	Pr> T	0.0001
Num ^= 0	4	Num > 0	4
M(Sign)	2	Pr>= M	0.1250
Sgn Rank	5	Pr>= S	0.1250
W:Normal	0.962634	Pr<W	0.7626
----- MUESTRA=4 -----			
Univariate Procedure			
Variable=VALOR			
Moments			
N	6	Sum Wgts	6
Mean	0.499333	Sum	2.996
Std Dev	0.010073	Variance	0.000101
Skewness	0.538444	Kurtosis	-0.24152
USS	1.49651	CSS	0.000507
CV	2.017303	Std Mean	0.004112
T:Mean=0	121.424	Pr> T	0.0001
Num ^= 0	6	Num > 0	6
M(Sign)	3	Pr>= M	0.0313
Sgn Rank	10.5	Pr>= S	0.0313
W:Normal	0.978903	Pr<W	0.9373
----- MUESTRA=5 -----			
Univariate Procedure			
Variable=VALOR			
Moments			
N	4	Sum Wgts	4
Mean	0.67725	Sum	2.709
Std Dev	0.050189	Variance	0.002519
Skewness	-1.83499	Kurtosis	3.498691
USS	1.842227	CSS	0.007557
CV	7.410677	Std Mean	0.025094
T:Mean=0	26.98809	Pr> T	0.0001
Num ^= 0	4	Num > 0	4
M(Sign)	2	Pr>= M	0.1250
Sgn Rank	5	Pr>= S	0.1250
W:Normal	0.779833	Pr<W	0.0690

### Independencia de los errores.-

Para completar el supuesto anterior se debe añadir que los errores están idéntica e independientemente distribuidos. De esta forma, si se ordenan los datos dentro de cualquier grupo en un orden lógico independiente de su magnitud (ese orden puede ser, por ejemplo, el mismo en el que se obtuvieron las medidas) se esperaría

que los errores se sucedan uno a otro en una secuencia aleatoria. En consecuencia, se supone que una gran sucesión de valores positivos, seguida por otra larga sucesión de valores negativos, es un hecho bastante improbable; y tampoco se debería esperar que los valores positivos y negativos puedan alternarse con regularidad.

La falta de independencia puede tener su origen en la correlación en el tiempo o en el espacio como consecuencia de que la asignación de los tratamientos a las unidades experimentales no ha sido correctamente aleatorizado. En un experimento se podría medir el efecto de un tratamiento mediante el registro de los pesos de diez individuos; la balanza puede sufrir una distorsión que dé como resultado subestimaciones sucesivas, compensadas por varias sobreestimaciones; o bien el operador de la balanza puede dar, en diferentes tiempos de pesada, sobreestimaciones y subestimaciones del verdadero peso. El problema de la no independencia de errores, como tantos otros problemas, puede ser resuelto con una aleatorización adecuada. Por ejemplo, se puede aleatorizar el orden en que se pesan los individuos de los diferentes grupos o tratamientos.

No existe una transformación o ajuste sencillo para evitar la falta de independencia de los errores. Hay que cambiar el diseño básico del experimento o la forma de realizarlo. Por ejemplo, en el experimento realizado con la balanza, posiblemente, parcial, se podría obtener la independencia de los errores volviendo a diseñar el experimento utilizando los diferentes tiempos de pesada como diferentes *bloques* o diferentes niveles de un nuevo factor.

Si los errores no son independientes, la validez de la prueba  $F$  puede estar gravemente invalidada.

El contraste de la aleatoriedad de una secuencia de mediciones de una variable (que se supone aleatoria) se puede realizar por las *pruebas de rachas* o de *ciclos* o bien por el *coeficiente de correlación serial*.

### Pruebas de rachas.-

Se va a estudiar, ahora, si determinados sucesos ocurren en sucesión aleatoria o, lo que es lo mismo, si la probabilidad de un acontecimiento dado es una función del resultado del acontecimiento anterior. Estas pruebas, que son de gran utilidad general, se conocen como *pruebas de ciclos* o *pruebas de rachas*. Para introducir este tipo de pruebas pongamos un ejemplo de fácil comprensión. Supóngase que se le dice a una persona que lance una moneda la aire 20 veces y anote el resultado de cada lanzamiento. Existe un gran número de formas diferentes en las que se podría dar este resultado. Una podría ser, siendo  $c$ =cara y  $x$ =cruz ésta

c c c c c c c c c c x x x x x x x x x x
---

Está claro que si se presenta este resultado, se podría afirmar que la secuencia no es aleatoria o, al menos, que es poco probable que ocurra una secuencia como esta.

El otro extremo sería la siguiente secuencia

C X C X C X C X C X C X C X C X C X C X

Esta alternancia regular entre ambos eventos haría sospechar que esta secuencia no es aleatoria o, al menos, que es poco probable.

Lo razonable, por probable, sería que hubiera *ciclos* o *racha* en que se alternaran caras con cruces, estos ciclos podrían ser de diferentes tamaños, desde uno hasta varias caras o cruces.

**Prueba de rachas para variables binomiales.-**

En el caso de que se tengan datos dicotómicos, la prueba sería la siguiente.

Supongamos que se tiene la secuencia

+ - + - + + + + - +

con  $n_1$  signos *más* y  $n_2$  signos *menos*. Se le llama racha o ciclo a cada cambio de signo; así en la secuencia anterior se han producido siete cambios de signo

$\frac{+}{1}$	$\frac{-}{2}$	$\frac{+}{3}$	$\frac{-}{4}$	$\frac{+}{5}$	$\frac{+}{6}$	$\frac{+}{7}$	$\frac{-}{8}$	$\frac{+}{9}$
---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------

Fijado el nivel de significación se compara el número de rachas ( $r$ ) con los valores críticos dados en la Tabla 11. Si  $r$  es menor o igual, o mayor o igual que los valores de la tabla se rechaza la hipótesis de independencia.

Si la muestra es grande ( $n_1$  ó  $n_2 > 20$ ) se puede realizar el contraste por medio de la siguiente aproximación normal

$$Z = \frac{r - \left( \frac{2 n_1 n_2}{n_1 + n_2} + 1 \right)}{\sqrt{\frac{2 n_1 n_2 (2 n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}$$

La tabla 11 se usa de la siguiente manera. Si se tiene una secuencia de, por ejemplo, 27 datos, con  $n_1=10$  y  $n_2=17$  (siempre se tiene que poner la secuencia más pequeña como  $n_1$ ) y con 12 rachas o cambios de signo, y se quiere hacer el contraste al  $\alpha=0.05$ , el valor crítico inferior es el que marca la columna de 0.025 y el superior el de la columna 0.975, que es 8 y 18, respectivamente, puesto que el número de rachas observadas esta entre los límites se acepta la hipótesis de aleatoriedad; si el número de rachas fuera igual o inferior a 8, o igual o superior a 18 se rechazaría la hipótesis de aleatoriedad. Si  $\alpha=0.01$ , habría que mirar en las columnas encabezadas por 0.005 y 0.995, etc.

### Ejemplo.-

En abejas y avispas, los huevos fertilizados dan como resultado hembras, mientras que los no fertilizados dan lugar a machos. Las hembras fertilizan algunos de estos huevos expulsando parte del esperma almacenado en las espermatecas durante la copulación, mientras que a otros huevos los deja sin fecundar. Con el fin de probar si esta operación es realizada por la hembra aleatoriamente o, por el contrario, sigue un orden de puesta, se tomaron los siguientes datos correspondientes a 18 descendientes de una avispa

H H H H H M M M H H H H H H M M M
-----------------------------------

se tiene, por tanto,  $n_1=6$  machos,  $n_2=12$  hembras y  $r=4$  rachas.

Los valores críticos para  $\alpha=0.05$  son

Límite inferior (en la columna 0.025) =  $r = 4$

Límite superior (en la columna 0.975) =  $r = 12$

Lo que indica que los huevos no fueron fecundados aleatoriamente.

Si se hubiera hecho la prueba por la aproximación a la normal, esta hubiera sido

$$Z = \frac{4 - \left(\frac{2 \times 6 \times 12}{6+12} + 1\right)}{\sqrt{\frac{2 \times 6 \times 12 (2 \times 6 \times 12 - 6 - 12)}{(6 + 12)^2 (6 + 12 - 1)}}} = -2.9124^{**}$$

$$Z_{(0.05)} = 1.64$$

$$Z_{(0.01)} = 2.33$$

Lo que indica que los huevos no fueron fecundados aleatoriamente.

### Pruebas de rachas para datos cuantitativos.-

Si lo que se tiene son datos cuantitativos y lo que se quiere saber es, como en el caso anterior, la independencia de los errores o falta de sesgo, existe otra prueba de rachas (*up and down*). Para aplicar la prueba se necesita convertir la variable observada en dicotómica: cero y uno, si y no, etc. Para ello se puede asignar el signo **más** a un valor cuando es mayor que el anterior y el signo **menos** si es menor que el anterior. De esta manera se tiene una secuencia de  $n-1$  signos + y signos - que permitirá contrastar la hipótesis de independencia.

Esta prueba se puede realizar por ajuste a la normal en el caso de que  $n > 10$  siendo los únicos parámetros necesarios el número de rachas y el número total de signos. Dicho ajuste es

$$Z = \frac{r - \frac{2n-1}{3}}{\sqrt{\frac{16n+29}{90}}}$$

**Ejemplo.-**

Se tiene el siguiente ANOVA de una vía

A	B	C
1452	1266	1275
1372	1080	1417
1273	1091	1186
1147	1302	989
1321	1514	1449
1083	1088	1564
1303	1218	1109
1245	1326	1299

Se elabora la secuencia de signos + y - según el criterio expuesto anteriormente

-	-	-	+	-	+	-	+	+	+	-	+	+	-	+	-	-	+	+	-	+
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16					

El primer signo, que es menos, es como consecuencia de que 1372 es inferior que 1452, el segundo signo, también menos, es porque 1273 es inferior que 1372, el tercer signo menos es porque 1147 es inferior que 1273, el cuarto signo, que es más es porque 1321 es mayor que 1147, etc.

El número total de signos es 23 y el número de rachas ( $r$ ) es de 16, la  $Z$  observada será

$$Z = \frac{16 - \frac{2 \times 23 - 1}{3}}{\sqrt{\frac{16 \times 23 + 29}{90}}} = 0.5152ns$$

$$Z_{(0.05/2)} = 1.64$$

por lo que se concluye que estos datos son independientes o aleatorios.

**Coefficiente de correlación serial.-**

Si se tienen  $n$  observaciones

$$X_1, X_2, \dots, X_n$$

se pueden formar los pares  $(X_i, X_{i+1})$ , de manera que se tendrá

$X_1 X_2$   
 $X_2 X_3$   
 $X_3 X_4$   
 ....

Con estas  $n-1$  parejas de valores se podrá calcular el coeficiente de correlación (ver Capítulo 15). Este coeficiente daría un valor más cercano a uno cuanto más se parezca un valor al anterior, y un valor cercano a cero cuanto más independiente sea un valor del anterior.

Este coeficiente se calcula de la siguiente manera

$$R_s = \frac{(n-1) \sum X_{i+1} X_i - \sum X_{i+1} \sum X_i}{\sqrt{[(n-1) \sum X_{i+1}^2 - (\sum X_{i+1})^2] [(n-1) \sum X_i^2 - (\sum X_i)^2]}}$$

Para valores de  $n$  mayores de 20 este coeficiente se distribuye aproximadamente como una normal con los siguientes parámetros

$$\mu = -\frac{1}{n-1}$$

$$\sigma^2 = \frac{n-2}{(n-1)^2}$$

Por lo que se puede realizar la prueba  $H_0: R_s=0$  con la siguiente  $Z_0$

$$Z_0 = \frac{(n-1)R_s + 1}{\sqrt{n-2}}$$

**Ejemplo.-**

Siguiendo con los mismos datos del ejemplo anterior, se pueden formar los pares siguientes

$X_i$	$X_{i+1}$	$X_i$	$X_{i+1}$	$X_i$	$X_{i+1}$
1452	1372	1266	1080	1275	1417
1372	1273	1080	1091	1417	1186
1273	1147	1091	1302	1186	989
1147	1321	1302	1514	989	1449
1321	1083	1514	1088	1449	1564
1083	1303	1088	1218	1564	1109
1303	1245	1218	1326	1109	1299
1245	1266	1326	1275		



$$n = 24$$

$$\sum X_{i+1} = 28917 \quad \sum X_{i+1}^2 = 36837877$$

$$\sum X_i = 29070 \quad \sum X_i^2 = 37258780$$

$$\sum X_{i+1}X_i = 36504174$$

$$R_s = \frac{(24-1)36504174 - 28917 \times 29070}{\sqrt{[(24-1)36837877 - 28917^2][(24-1)37258780 - 29070^2]}} = -0.0890$$

$$Z_o = \frac{(24-1) - 0.088989 + 1}{\sqrt{24-2}} = -0.2232ns$$

$$Z_{005} = 1.96$$

Por lo que se concluye que estos datos son independientes o aleatorios.

### Archivo del programa SAS (C6-2.SAS)-

```

title 'Coeficiente de correlación serial';
options ls=75 ps=60;
data corseri;
infile 'c6-2.dat';
input val @@;
data corse2;
set corseri (firstobs=2);
xii = val;
sumn + 1;
sumxii + val;
sumxii2 + val*val;
run;
data corse3;
set corseri (obs=23);
xi = val;
sumn + 1;
sumxi + val;
sumxi2 + val*val;
run;
data corse4;
merge corse2 corse3;
by sumn;
sumxiixi + xii*xi;
run;
data corse5;
set corse4 (firstobs=23);
nrs = sumn*sumxiixi - sumxii*sumxi;
drs = sqrt((sumn*sumxi2 - sumxii**2)*(sumn*sumxi2 - sumxi**2));
rs = nrs/drs;
zo = (sumn*rs+1)/sqrt(sumn-1);
if 1-zo>1 then zo2=zo;
if 1-zo<1 then zo2=-zo;
prz=probnorm(zo2)*2;
proc print;
var rs zo prz ;
run;

```

## Archivo de datos (C6-2.DAT).-

1452	1372	1273	1147	1321	1083	1303	1245
1266	1080	1091	1303	1514	1088	1218	1326
1275	1417	1186	989	1449	1564	1109	1299

## Archivo de resultados (C6-2.LST) -

Coeficiente de correlación serial			
OBS	RS	ZO	PRZ
1	-0.088805	-0.22227	0.82411

## Homogeneidad de varianzas.-

La homogeneidad de varianzas entre un grupo de submuestras es una precondition importante de muchas pruebas estadísticas. El término estadístico para denominar esta homogeneidad es el de *homoscedasticidad*, mientras que el inverso es el de *heteroscedasticidad*.

Si se tiene, por ejemplo, un ANOVA de una vía, el supuesto es que los datos de los diferentes tratamientos son submuestras de una misma población donde el parámetro que ha modificado los tratamientos es  $\mu$ , por lo que se tendrá  $t$  distribuciones con, posiblemente, diferente  $\mu$  y la misma  $\sigma^2$ . Por lo que se supone que cada varianza muestral es una estima de la misma varianza paramétrica, por lo que la hipótesis de homogeneidad de varianzas es obvia.

Las razones por las que se pueden presentar varianzas heterogéneas en un diseño puede ser varias.

Puede ocurrir que algunas de las poblaciones sean inherentemente más variables que otras. Algunas especies son relativamente uniformes para un carácter, mientras que otras son muy variables para el mismo carácter.

En un ANOVA de los resultados de un experimento puede ocurrir que una muestra se haya obtenido en condiciones menos normalizadas que otras y, por lo tanto, posea una varianza mayor.

Existen también muchos casos en los que la heterogeneidad de las varianzas es consecuencia de haber elegido mal la escala de medida. En algunas escalas de medidas, las varianzas varían en función de la media, de manera que las diferencias entre medias provocan varianzas heterogéneas. Por ejemplo, en las variables que siguen una distribución de Poisson, la varianza es, de hecho, igual que la media, por lo que las poblaciones con mayor media tendrán mayor varianza. Estas desviaciones de la hipótesis de *homoscedasticidad* pueden, a menudo, corregirse mediante una transformación, como se verá más adelante.

Las consecuencias de una heterogeneidad moderada de las varianzas no son demasiado graves para la prueba  $F$ , aunque las comparaciones con un único grado de

libertad pueden alejarse mucho de la exactitud. En todo caso, la prueba  $F$  no es robusta para este supuesto, aunque si lo es para el supuesto de normalidad.

Ya se vio en el Capítulo 4 la prueba  $F$  para la homogeneidad de dos varianzas. Un método aproximado para probar la homogeneidad de más de dos varianzas puede ser algo semejante a esta prueba  $F$ , dividiendo la varianza mayor por la menor; si esta prueba es *no significativa* es porque las varianzas son homogéneas, mientras que si es significativa no se puede concluir que sean heterogéneas, por lo que hay que realizar necesariamente la prueba de *Bartlett*.

### Prueba de *Bartlett*-

Es la prueba clásica para contrastar la hipótesis de igualdad de  $t$  varianzas. El mayor inconveniente de esta prueba es el ser muy sensible a la no normalidad, por lo que un resultado significativo puede, por tanto, indicar tanto no normalidad como heteroscedasticidad.

El método de *Bartlett* es como sigue

Sean  $t$  muestras.

Con varianza muestral,  $S^2_1, S^2_2, \dots, S^2_t$ ,

Y con  $g_1, g_2, \dots, g_t$ .

Se calcula la varianza ponderada

$$\bar{S}^2 = \frac{\sum_i g_i S_i^2}{\sum_i g_i}$$

Se calcula la expresión

$$B = GL \ln S^2 - \sum_i (g_i \ln S_i^2)$$

siendo  $GL = \sum g_i$ .

Si  $B$  es menor que el valor de  $\chi^2_{(t-1)}$  se acepta la hipótesis de homogeneidad de varianzas. Si es mucho mayor que el valor de las tablas se rechaza la hipótesis nula de homogeneidad de varianzas. Si es ligeramente mayor que el valor de  $\chi^2_{(t-1)}$  se tiene que dividir el valor de  $B$  por el coeficiente de corrección,  $C$ , siguiente

$$C = 1 + \frac{\sum_i \frac{1}{g_i} - \frac{1}{GL}}{3(t-1)}$$

El valor  $B/C$  se distribuye como una  $\chi^2_{(t-1)}$ .

**Ejemplo.-**

Comprobar si existe uniformidad de varianzas en los datos de los dos ejemplos anteriores. Dicha experiencia era un diseño completamente aleatorio con tres tratamientos.

$t_1$	$t_2$	$t_3$
1452	1266	1275
1372	1080	1417
1273	1091	1186
1147	1302	989
1321	1514	1449
1083	1088	1564
1303	1218	1109
1245	1326	1299

$$\begin{aligned}
 S_1^2 &= 13969.7142 & S_2^2 &= 22603.9821 & S_3^2 &= 35834.5714 \\
 \ln S_1^2 &= 9.544647 & \ln S_2^2 &= 10.0258814 & \ln S_3^2 &= 10.4866684 \\
 gl_1 &= 7 & gl_2 &= 7 & gl_3 &= 7 \\
 \bar{S}^2 &= 24136.0892 \\
 GL &= 21 \\
 \ln \bar{S}^2 &= 10.0914635
 \end{aligned}$$

$$B = 21 \times 10.0914635 - 210.4003773 = 1.5203568ns$$

$$\chi^2_{(2; 0.05)} = 5.992$$

Como  $B$  es inferior que el valor de  $\chi^2_{(2; 0.05)}$  se acepta la hipótesis nula de homogeneidad de varianzas.

Si se hubiese probado por el método aproximado de la  $F$ , el resultado hubiera sido

$$\begin{aligned}
 F &= \frac{35834.5714}{13969.7142} = 2.5652ns \\
 F_{(7,7; 0.05)} &= 4.99
 \end{aligned}$$

Se hubiera, por lo tanto, concluido lo mismo, es decir, que las varianzas son homogéneas.

### Archivo del programa SAS (C6-3.SAS).-

```
title 'Prueba de Bartlett';
option ls=75 ps=60;
data cedas1;
infile 'c6-3.dat';
muestras=3;
  do t=1 to muestras;
    input val @@ ; output;
  end;
proc means noprint n var;
class t;
var val;
output out=varianza n=n var=var t;
data cedas2;
set varianza (firstobs=2);
sumn + (n-1);
sumin + 1/(n-1);
sumvar + var*(n-1);
sumlnvar + (n-1)*log(var);
s2=sumvar/sumn;
ji2=sumn*log(s2)-sumlnvar;
if t=1 then t=2;
c=1+(sumin-(1/sumn))/(3*(t-1));
ji2adj=ji2/c;
prji2adj=1-probchi(ji2adj, (t-1));
prji2=1-probchi(ji2, (t-1));
run;
proc print;
var s2 ji2 prji2 ji2adj prji2adj;
run;
```

### Archivo de datos (C6-3.DAT).-

1452	1266	1275
1372	1080	1417
1273	1091	1186
1147	1302	989
1321	1514	1449
1083	1088	1564
1303	1218	1109
1245	1326	1299

### Archivo de resultados (C6-3.LST).-

Prueba de Barlett					
OBS	S2	JI2	PRJI2	JI2ADJ	PRJI2ADJ
1	13969.71	-0.00000	.	-0.00000	.
2	18286.85	0.40143	0.52635	0.37466	0.54047
3	24136.09	1.52036	0.46758	1.42959	0.48929

Como siempre, en programas como estos, se mira solo la última línea.

## Transformaciones.-

La finalidad de transformar las observaciones es la de corregir algunas de las desviaciones originadas por el no cumplimiento de algunas de las condiciones paramétricas o supuestos del análisis de varianza. Un hecho afortunado de las transformaciones es que, muy a menudo, varias desviaciones de los supuestos son subsanadas simultáneamente mediante la misma transformación. De manera que haciendo los datos homoscedásticos, también se aproximan a la normalidad y se produce, asimismo, la aditividad de los efectos de los tratamientos.

Cuando se aplica una transformación, las pruebas de significación se realizan sobre los datos transformados, aunque los valores calculados de las medias se dan, generalmente, en la escala no transformada.

Las transformaciones más comunes son

*Logarítmica*, para cuando los efectos son multiplicativos (no aditivos) o cuando existen relaciones alométricas entre los datos.

*Raíz cuadrada*, para variable de *Poisson*.

*Angular* o *arcoseno*, para variables binomiales.

*Probit*, también para variables binomiales, especialmente con datos de mortalidad.

Las tres primeras se estudiarán a continuación, la *logarítmica* se volverá a estudiar en el Capítulo 14, y la transformación *probit* solo se verá en el Capítulo 12.

## Transformación logarítmica.-

La transformación más común es la conversión de todos los datos en logaritmos, generalmente en logaritmos decimales. Siempre que la media esté positivamente relacionada con la varianza (a mayores medias corresponden mayores varianzas) la transformación logarítmica es muy probable que solucione el problema y haga la varianza independiente de la media. Esta transformación corrige la no normalidad y la no aditividad del modelo.

Algunas de las precauciones a tener con esta transformación es que no puede aplicarse a datos negativos, si existen ceros entre los datos, para evitar los infinitos, se recomienda sumar la unidad a todos los datos, y si existen datos decimales, para evitar datos transformados negativos, se multiplican por 10, o cualquier múltiplo de 10 que elimine los decimales, puesto que esto no produce ningún efecto sobre los análisis subsecuentes.

### Ejemplo.-

Se tienen los pesos en libras de animales de tres especies tratados con un nuevo producto vitamínico y animales de control (sin vitamina), en un experimento de bloques aleatorios. Los datos son los siguientes

Especie	Tratamiento	Bloque			
		I	II	III	IV
Ratones	Control	0.18	0.30	0.28	0.44
	Vitamina	0.32	0.40	0.42	0.46
Gallina	Control	2.0	3.0	1.8	2.8
	Vitamina	2.5	3.3	2.5	3.3
Oveja	Control	108.0	140.0	135.0	165.0
	Vitamina	127.0	153.0	148.0	176.0

### Archivo del programa SAS (C6-4.SAS).-

```
title 'Transformación raíz logarítmica';
options ls=75 ps=60;
data transfor;
infile 'c6-4.dat';
do bloque=1 to 4;
  input especie $ vitamina $ peso @@ ;
  logpeso=log10(peso*10);
  output;
end;
proc anova;
class bloque especie vitamina;
model peso = bloque especie vitamina especie*vitamina;
proc anova;
class bloque especie vitamina;
model logpeso = bloque especie vitamina especie*vitamina;
run;
```

### Archivo de datos (C6-4.DAT).-

```
rat con 0.18 rat con 0.30 rat con 0.28 rat con 0.44
rat vit 0.32 rat vit 0.40 rat vit 0.42 rat vit 0.46
gal con 2.00 gal con 3.00 gal con 1.80 gal con 2.80
gal vit 2.50 gal vit 3.30 gal vit 2.50 gal vit 3.30
ove con 108.00 ove con 140.00 ove con 135.00 ove con 165.00
ove vit 127.00 ove vit 153.00 ove vit 148.00 ove vit 176.00
```

Archivo de resultados (C6-4.LST)-

Transformación raíz logarítmica						1
Dependent Variable: PESO						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	8	109697.680	13712.210	110.01	0.0001	
Error	15	1869.725	124.648			
Corrected Total	23	111567.405				
	R-Square	C.V.	Root MSE	PESO Mean		
	0.983241	22.78490	11.1646	49.0000		
Source	DF	Anova SS	Mean Square	F Value	Pr > F	
BLOQUE	3	984.000	328.000	2.63	0.0881	
ESPECIE	2	108321.160	54160.580	434.51	0.0001	
VITAMINA	1	142.107	142.107	1.14	0.3025	
ESPECIE*VITAMINA	2	250.413	125.207	1.00	0.3896	
Dependent Variable: LOGPESO						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	8	28.7528117	3.5941015	1167.12	0.0001	
Error	15	0.0461918	0.0030795			
Corrected Total	23	28.7990035				
	R-Square	C.V.	Root MSE	LOGPESO Mean		
	0.998396	3.267379	0.05549	1.69839		
Source	DF	Anova SS	Mean Square	F Value	Pr > F	
BLOQUE	3	0.1204960	0.0401653	13.04	0.0002	
ESPECIE	2	28.5722445	14.2861223	4639.17	0.0001	
VITAMINA	1	0.0503652	0.0503652	16.36	0.0011	
ESPECIE*VITAMINA	2	0.0097060	0.0048530	1.58	0.2392	

Como se puede observar, cuando se realiza el análisis de varianza con los datos originales, da significativo sólo la diferencia de peso entre especies, lo cual era de esperar. Sin embargo, puede parecer muy extraño que no se encuentren diferencias significativas debidas a la vitamina, especialmente porque cada animal que recibió vitamina mostró en cada repetición un mayor peso que el animal de control correspondiente. Parece también extraño que no se encuentre evidencia de interacción entre los efectos de la vitamina y la especie. Si se acepta este análisis como bueno, se tendría que admitir que el experimento es un fracaso. Lo único que concluimos es que los ratones, las gallinas y las ovejas pesan diferente.

Sin embargo, cuando se realiza el análisis de varianza con la transformación logarítmica de los datos se obtienen conclusiones más satisfactorias, como es la diferencia significativa de peso entre los animales tratados con vitamina y los del control.



## Transformación de la raíz cuadrada.-

Cuando los datos son números enteros como pueden ser los conteos de, por ejemplo, los insectos que puede haber en una hoja o las células sanguíneas en un hematocitómetro, es conveniente trabajar con la raíz cuadrada de los datos. Recuérdese que la distribución de tales tipos de datos es la de *Poisson* y no la normal, y que en una distribución de *Poisson* la varianza es igual a la media. Por tanto, la varianza y la media no sólo no serán independientes, sino que variarán idénticamente.

Cuando se analizan datos de este tipo, se está violando los diferentes supuestos, descritos anteriormente, que debe cumplir un análisis de varianza. Los errores no están normalmente distribuidos y las varianzas están relacionadas con las medias. La transformación de los datos en sus raíces cuadradas generalmente hará que las varianzas sean independientes de las medias. Si los datos incluyen el cero entre sus valores, o si el valor máximo es inferior a diez, se ha comprobado la conveniencia de codificar todos los valores añadiéndole 0.5 antes de extraer la raíz cuadrada. La transformación, por tanto, será

$$\sqrt{X + \frac{1}{2}}$$

Se ha demostrado que también da buenos resultados la transformación

$$\sqrt{X + \frac{3}{8}}$$

En general, se puede decir que los datos que requieren la transformación raíz cuadrada no violan, tan drásticamente, los supuestos del análisis de varianza como los datos que requieren una transformación logarítmica. Consecuentemente, los cambios causados, en el análisis, por esta transformación no son tan espectaculares.

### Ejemplo.-

Se quiere probar la efectividad de diez insecticidas, para ello se realiza un diseño con cuatro bloques aleatorios y tomando un nivel más del tratamiento insecticida como control. Los datos son número de insectos vivos contabilizados, por unidad experimental, transcurrido un tiempo prudencial después de la aplicación del insecticida

<i>Insecticida</i>	<i>Bloque</i>			
	I	II	III	IV
1	7	5	4	1
2	6	1	2	1
3	6	2	1	0
4	0	1	2	0
5	1	0	1	2
6	5	14	9	15
7	8	6	3	6
8	3	0	5	9
9	4	10	13	5
10	6	11	5	2
Control	8	11	2	6

### Archivo del programa SAS (C6-5.SAS).-

```

title 'Transformación raíz raíz cuadrada';
options ls=75 ps=60;
data traraiz;
infile 'c6-5.dat';
do bloque=1 to 4;
  do insectic=1 to 11;
    input insectos @@;
    raizinse=sqrt(insectos+0.5);
    output;
  end;
end;
proc anova;
class insectic bloque;
model insectos = insectic bloque;
means insectic/duncan;
proc anova;
class insectic bloque;
model raizinse = insectic bloque;
means insectic/duncan;
run;

```

### Archivo de datos (C6-5.DAT).-

7	6	6	0	1	5	8	3	4	6	8
5	1	2	1	0	14	6	0	10	11	11
4	2	1	2	1	9	3	5	13	5	2
1	1	0	0	2	15	6	9	5	2	6

## Archivo de resultados (C6-5.LST).-

Transformación raíz raíz cuadrada										
Dependent Variable: INSECTOS										
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F					
Model	13	392.250000	30.173077	2.94	0.0072					
Error	30	308.000000	10.266667							
Corrected Total	43	700.250000								
	R-Square	C.V.	Root MSE	INSECTOS Mean						
	0.560157	67.45608	3.20416	4.75000						
Source	DF	Anova SS	Mean Square	F Value	Pr > F					
INSECTIC	10	380.000000	38.000000	3.70	0.0026					
BLOQUE	3	12.250000	4.083333	0.40	0.7556					
Duncan's Multiple Range Test for variable: INSECTOS										
NOTE: This test controls the type I comparisonwise error rate, not the experimentwise error rate										
Alpha= 0.05 df= 30 MSE= 10.26667										
Number of Means	2	3	4	5	6	7	8	9	10	11
Critical Range	4.627	4.863	5.015	5.124	5.207	5.271	5.323	5.365	5.400	5.430
Means with the same letter are not significantly different.										
Duncan Grouping										
			Mean	N	INSECTIC					
		A	10.750	4	6					
B		A	8.000	4	9					
B		A C	6.750	4	11					
B	D	A C	6.000	4	10					
B	D	A C	5.750	4	7					
B	D	C	4.250	4	8					
B	D	C	4.250	4	1					
	D	C	2.500	4	2					
	D	C	2.250	4	3					
	D		1.000	4	5					
	D		0.750	4	4					
Analysis of Variance Procedure										
Dependent Variable: RAZINSE										
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F					
Model	13	20.5250759	1.5788520	3.19	0.0042					
Error	30	14.8407826	0.4946928							
Corrected Total	43	35.3658585								
	R-Square	C.V.	Root MSE	RAZINSE Mean						
	0.580364	33.35583	0.70334	2.10861						
Source	DF	Anova SS	Mean Square	F Value	Pr > F					
INSECTIC	10	19.9935278	1.9993528	4.04	0.0014					
BLOQUE	3	0.5315481	0.1771827	0.36	0.7836					
Duncan's Multiple Range Test for variable: RAZINSE										
NOTE: This test controls the type I comparisonwise error rate, not the experimentwise error rate										
Alpha= 0.05 df= 30 MSE= 0.494693										

Number of Means	2	3	4	5	6	7	8	9	10	11
Critical Range	1.016	1.067	1.101	1.125	1.143	1.157	1.168	1.178	1.185	1.192
Means with the same letter are not significantly different.										
Duncan Grouping					Mean	N	INSECTIC			
		A			3.2931	4	6			
B		A			2.8453	4	9			
B		A	C		2.6093	4	11			
B		A	C		2.4713	4	7			
B		A	C		2.4668	4	10			
B		D	C		2.1075	4	1			
B		D	C		2.0013	4	8			
		D	C		1.6450	4	2			
		D	C		1.5156	4	3			
		D			1.1844	4	5			
		D			1.0550	4	4			

Como se puede observar, los resultados de los dos análisis no son demasiado diferentes, puesto que ambos muestran un efecto de los insecticidas altamente significativo. De todas maneras es útil resaltar que el valor de la *F* es mayor para los datos transformados, lo que podría ser importante si hubiera estado cerca de los valores críticos. Sin embargo sí se registran algunas diferencias importantes en la separación de medias cuando se hace un contraste múltiple de medias, como es el de *Duncan*.

Notemos que en los datos transformados el insecticida 7 y 10 no son estadísticamente iguales al 5 y al 4 (que son los mejores) mientras que en los datos en bruto si son iguales. El efecto general de la transformación es incrementar la precisión con la cual se puede medir las diferencias entre medias pequeñas. Esto es muy deseable en experiencias como la presente, ya que no se está tan interesados en las diferencias entre dos tratamientos relativamente ineficaces como en detectar pequeñas, pero significativas diferencias entre insecticidas eficaces que permitan un mejor control.

**Transformación angular o arcoseno.-**

Otro tipo de datos que pueden requerir transformación son los basados en conteos expresados como porcentajes o proporciones de la muestra total. Por regla general, tales datos tienen una distribución *binomial*, en vez de la distribución normal que requiere el análisis de varianza. Recuérdese que la varianza de la distribución binomial es  $\sigma^2=npq$ , y dado que  $\mu = np$  y que  $q=(1-p)$ , está claro que en una distribución binomial la varianza será función de la media, pero no como en los casos anteriores en que medias grandes tienden a tener varianzas grandes, sino que, en los datos binomiales, las varianzas tienden a ser pequeñas en los dos extremos de los rangos de valores (cercanos al 0 y al 1), y mayores en el medio (alrededor del 0.5). La transformación *arcsen* impide esta posibilidad.

Esta transformación consiste en sustituir cada dato o proporción por el ángulo cuyo seno es la raíz cuadrada de la proporción (no porcentaje).

$$\Theta = arcsen(\sqrt{p})$$

donde  $p$  es una proporción.

El término *arcsen* es sinónimo de la función  $\text{sen}^{-1}$ , la cual representa el ángulo (en grados) cuyo seno es la raíz cuadrada del dato.

La transformación arcoseno alarga ambas colas de la distribución de proporciones y comprime la parte central, aproximándola a una distribución normal.

La varianza esperada de tal distribución de arcoseno es

$$\sigma^2 = \frac{180}{4\pi^2 n} = \frac{820.8}{n}$$

Cuando  $\Theta$  se mide en radianes en lugar de grados, entonces

$$\sigma^2 = \frac{1}{4n}$$

No es necesario realizar esta transformación si el rango de proporciones está entre 0.3 y 0.7, puesto que en ese rango la distribución se aproxima bastante a una normal.

Si se tiene valores de  $p=0$ , se aconseja sean sustituidos por

$$\frac{1}{2n} \quad \text{ó} \quad \frac{1}{4n}$$

Los valores  $p=1$ , se aconseja sustituir por

$$1 - \frac{1}{2n} \quad \text{ó} \quad 1 - \frac{1}{4n}$$

Si el tamaño de muestra es grande se puede calcular las proporciones como siempre ( $p = X/n$ ), pero si  $n$  es pequeño se aconseja calcular las proporciones, antes de transformarlas, de la siguiente manera

$$p = \frac{X + \frac{3}{8}}{n + \frac{3}{4}}$$

### Ejemplo.-

Se tiene el porcentaje de plantas de soja infectadas con el llamado tumor de *Hudson* que provoca abultamientos en los tallos. Los datos se tomaron de seis variedades en cuatro bloques

	Bloque		Variedad			
	1	2	3	4	5	6
1	19.3	10.1	25.2	14.0	3.3	3.1
2	29.2	34.7	36.5	30.2	35.8	9.6
3	1.0	14.0	23.4	7.2	1.1	1.0
4	6.4	5.6	12.9	8.9	2.0	1.0

### Archivo del programa SAS (C6-6.SAS).-

```

title 'Transformación Arcoseno';
options ls=75 ps=60;
data transang;
infile 'c6-6.dat';
do bloque=1 to 4;
  do variedad=1 to 6;
    input pif @@ ;
    ar=arsin(sqrt(pif/100));tapif=ar*360/2/3.1415927;output;
  end;
end;
proc anova;
class variedad bloque;
model pif = variedad bloque;
means variedad / snk;
run;
proc anova;
class variedad bloque;
model tapif = variedad bloque;
means variedad / snk;
run;

```

### Archivo de datos (C6-6.DAT).-

19.3	10.1	25.2	14.0	3.3	3.1
29.2	34.7	36.5	30.2	35.8	9.6
1.0	14.0	23.4	7.2	1.1	1.0
6.4	5.6	12.9	8.9	2.0	1.0

### Archivo de resultados (C6-6.LST).-

Transformación Arcoseno						
Dependent Variable: PIF						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	8	2952.36500	369.04563	11.53	0.0001	
Error	15	480.03458	32.00231			
Corrected Total	23	3432.39958				
	R-Square	C.V.	Root MSE	PIF Mean		
	0.860146	40.46778	5.65706	13.9792		
Source	DF	Anova SS	Mean Square	F Value	Pr > F	
VARIEDAD	5	937.28708	187.45742	5.86	0.0034	
BLOQUE	3	2015.07792	671.69264	20.99	0.0001	

Student-Newman-Keuls test for variable: PIF  
 NOTE: This test controls the type I experimentwise error rate under the complete null hypothesis but not under partial null hypotheses.

Alpha= 0.05 df= 15 MSE= 32.00231  
 Number of Means 2 3 4 5 6  
 Critical Range 8.5261058 10.390256 11.529011 12.352144 12.996342

Means with the same letter are not significantly different.

SNK Grouping	Mean	N	VARIEDAD
A	24.500	4	3
B A	16.100	4	2
B A C	15.075	4	4
B A C	13.975	4	1
B C	10.550	4	5
C	3.675	4	6

Dependent Variable: TAPIF

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	2274.40481	284.30060	11.78	0.0001
Error	15	362.08997	24.13933		
Corrected Total	23	2636.49478			

R-Square 0.862662 C.V. 24.77286 Root MSE 4.91318 TAPIF Mean 19.8329

Source	DF	Anova SS	Mean Square	F Value	Pr > F
VARIEDAD	5	883.74548	176.74910	7.32	0.0012
BLOQUE	3	1390.65934	463.55311	19.20	0.0001

Student-Newman-Keuls test for variable: TAPIF  
 NOTE: This test controls the type I experimentwise error rate under the complete null hypothesis but not under partial null hypotheses.

Alpha= 0.05 df= 15 MSE= 24.13933  
 Number of Means 2 3 4 5 6  
 Critical Range 7.4049597 9.0239823 10.012996 10.72789 11.287379

Means with the same letter are not significantly different.

SNK Grouping	Mean	N	VARIEDAD
A	29.320	4	3
B A	22.571	4	2
B A	22.058	4	4
B A	19.791	4	1
B C	15.342	4	5
C	9.917	4	6

Como se puede observar, y como ocurría con la transformación raíz cuadrada, los resultados de los dos análisis no son demasiado diferentes, puesto que ambos muestran un efecto significativo de la variedades. De todas maneras es útil resaltar que el valor de la  $F$  es mayor para los datos transformados, lo que podría ser importante si hubiera estado cerca de los valores críticos. Sin embargo sí se registran una separación más nítida cuando se hace un contraste múltiple de medias, como es el de  $SNK$ .

Tal vez en un experimento con más datos se note más la influencia de esta transformación.

### **Bibliografía:**

- Cochran, WG y Cox, GM.* 1965. DISEÑOS EXPERIMENTALES. Ed. TRILLAS. México.
- Li, Ching Chun.* 1969. INTRODUCCIÓN A LA ESTADÍSTICA EXPERIMENTAL. Ed. Omega. Barcelona.
- Lite, TM, y Jackson Hills, F.* 1987. MÉTODOS ESTADÍSTICOS PARA LA INVESTIGACIÓN EN LA AGRICULTURA. Ed. TRILLAS. México.
- Littel, R.C., Freund, R.J. and Spector, P.C.* 1991. SAS® FOR LINEAR MODELS. SAS Institute Inc., Cary, NC, USA.
- Ostle, B.* 1965. ESTADÍSTICA APLICADA. Ed. Limusa-Wiley. México.
- Ruiz-Maya, L.* 1983. MÉTODOS ESTADÍSTICOS DE INVESTIGACIÓN. Ed. INE. Madrid.
- Sokal, R.R. y Rohlf, F.J.* 1994. BIOMETRY. Ed. W.H.FREEMAN. San Francisco.
- Snedecor, G.W. y Cochran, W.G.* 1971. MÉTODOS ESTADÍSTICOS. Ed. C.E.C.S.A. México.
- Steel, R.* 1996. PRINCIPLES AND PROCEDURES OF STATISTICS. Ed. McGRAW-HILL Education. New York .
- SAS Institute Inc.* 1990. SAS/STAT USER'S GUIDE. Volume 1 and 2. Cary, NC, USA.





## **CAPÍTULO 7**

# **Modelos Factoriales**



# CAPÍTULO 7

## Modelos Factoriales

### Número de medidas por casilla.-

Cuando se tenga un modelo lineal con dos factores no siempre se va a tener una sola medida por casilla tal como es el caso de los últimos ejemplos del Capítulo 5.

Ocurre a menudo que se tienen varias medidas por casilla como consecuencia de que se tienen más de  $t$  unidades experimentales por bloque. En el último ejemplo del capítulo 5, se puede tener más de tres animales por raza para realizar la experiencia; por ejemplo, supóngase que se ha realizado la experiencia en seis animales por raza, por lo que cada tratamiento se repite dos veces en cada raza.

También puede ocurrir que se tenga más de una medida por casilla porque, aunque cada casilla constituye una unidad experimental, se han realizado varias medidas por unidad experimental, es decir, no coincide la *unidad experimental* con la *unidad de muestreo* o de medida; por ejemplo, supóngase que en el ejemplo anterior los tratamientos fueran tres desparasitadores probados en cuatro rebaños, cada uno en cada uno de tres animales elegidos al azar dentro de cada rebaño, pero en lugar de realizar una sola medida en cada animal se realiza dos medidas en sendas partes del cuerpo, es decir, se toman dos muestras de cada unidad experimental, con lo que en cada casilla se tendrá dos medidas correspondientes a dos *unidades de muestra* pero de una sola *unidad experimental*, que es cada animal. En ese caso se tendrá dos errores el *error experimental* o debido a la variabilidad individual de las diferentes unidades experimentales, y el *error de muestreo* debido a la variabilidad individual de las diferentes unidades de muestreo o medida. Este ejemplo no es realmente un ejemplo de varias repeticiones por casilla, sino que en parte es como los descritos en el epígrafe anterior y en parte es como lo que se describirá en el próximo epígrafe, es decir, es como un caso de una sola medida por casilla, pero teniendo dentro de cada casilla medidas *jerarquizadas* (ver más adelante) a la unidad experimental de cada casilla. Como no se va a volver a ver este caso diremos que los contrastes de los dos factores se hace con la interacción (ver más adelante).

**Ejemplo.-**

Para continuar la explicación centrémonos en el autentico caso de repeticiones dentro de las casillas y para ello volvamos al primer ejemplo que hemos estado utilizando a través del capítulo 5.

Supóngase que en este caso se nos hubiera informado que los animales utilizados eran, porque así interesaba, de dos razas diferentes. Para recapitular y ordenar correctamente las ideas, recuérdese que el primer factor se trataba de tres *tratamientos* que podían ser tres dietas o tres antibióticos o tres tóxicos o tres métodos de purificación, etc., es decir, un factor con tres únicos niveles del cual queremos saber si existe diferencias entres dichos niveles y en caso positivo, saber cuál o cuales es el mejor. Con respecto al segundo factor, *raza*, ocurre otro tanto, son dos razas y queremos saber si son diferentes y en caso positivo saber cual es la mejor. Los datos son los mismos del ejemplo anterior y del primer ejemplo pero ordenados de la siguiente manera.

	$T_1$	$T_2$	$T_3$			
$R_1$	$X_{111}= 5$	$X_{211}= 6$	$X_{311}=4$	$X_{.1}=29$	$\bar{X}_{.1}=4.8$	$\rho= 1.8$
	$X_{112}= 4$	$X_{212}= 7$	$X_{312}=3$			
$R_2$	$X_{121}= 1$	$X_{221}= 2$	$X_{321}=1$	$X_{.2}= 7$	$\bar{X}_{.2}=1.2$	$\rho=-1.8$
	$X_{122}= 2$	$X_{222}= 1$	$X_{322}=0$			
	$X_{1..}=12$	$X_{2..}=16$	$X_{3..}=8$	$X_{...}=36$		
	$\bar{X}_{1..}= 3$	$\bar{X}_{2..}= 4$	$\bar{X}_{3..}=2$	$\bar{X}_{...}= 3$		
	$\tau_1 = 0$	$\tau_2 = 1$	$\tau_3 =-1$			

Como en el ejemplo anterior, el que una unidad experimental pertenezca a una raza o a otra no es aleatorio, sino que es el experimentador el que lo determina con arreglo a criterios ajenos al problema; normalmente esta segunda fuente de variación es debida a factores ambientales o a criterios de clasificación. Pero dentro de cada uno de los niveles de este segundo factor los tratamientos se aplican aleatoriamente. En este ejemplo concreto, se han tenido que elegir aleatoriamente seis animales de cada raza para aplicarle aleatoriamente uno de los tres tratamientos a cada dos. Por lo tanto, el que un animal pertenezca a una raza o a otra no es aleatorio mientras que los tratamientos han sido aplicados aleatoriamente a cada animal dentro de cada raza.

Tal como se ha hecho con los ejemplos anteriores, la relación existe entre estos valores puede expresarse por medio de desviaciones, de manera que la desviación total de una unidad experimental (valor de dicha unidad menos la media general) es igual a la desviación del grupo en que está dicho individuo (media del grupo menos media general) más la desviación debida a la variación individual (lo que el individuo tiene al final con respecto a lo que tendría por el grupo al que pertenece). Pero en este caso, como en el anterior, cada individuo pertenece a dos grupos, uno con respecto a  $T_i$  y otro con respecto a  $R_j$ , por lo que la desviación total será igual a la  $i$ -ésima desviación del grupo  $T_i$  más la  $j$ -ésima desviación de grupo  $R_j$  más la desviación individual, es decir

$$(X_{ijk} - \bar{X}_{...}) = (\bar{X}_{i..} - \bar{X}_{...}) + (\bar{X}_{.j.} - \bar{X}_{...}) + (X_{ijk} - \bar{X}_{ij.})$$

Al contrario de lo que ocurría en el ejemplo anterior, aquí hay más de una medida dentro de cada casilla, por lo que el último sumando de esta expresión no es cero y, por tanto, puede medirse la desviación individual o debida al *error*. Por lo que la desviación total de una unidad experimental es igual a la suma de las desviaciones debidas a los dos grupos a los que pertenece dicha unidad experimental más la desviación individual

Pasando la media general del termino de la derecha al de la izquierda, se obtiene

$$X_{ijk} = \bar{X}_{...} + (\bar{X}_{i..} - \bar{X}_{...}) + (\bar{X}_{.j.} - \bar{X}_{...}) + (X_{ijk} - \bar{X}_{ij.})$$

Lo que quiere decir que la medida de una unidad experimental es igual a la suma de cuatro sumandos (por el momento, igual que en el anterior modelo), el primero es la estima de la *media general*,  $\mu$ , el segundo es la estima del efecto del *tratamiento*,  $\tau_i$ , el tercero es la estima del efecto de la *raza*,  $\rho_j$ , y el cuarto es la estima del *error experimental*,  $\varepsilon_{ij}$ . Por lo que en este caso, el modelo lineal es el mismo del ejemplo, este es

$$X_{ijk} = \bar{X}_{...} + (\bar{X}_{i..} - \bar{X}_{...}) + (\bar{X}_{.j.} - \bar{X}_{...}) + (X_{ijk} - \bar{X}_{ij.})$$

$$X_{ijk} = \mu + T_i + R_j + \varepsilon_{ijk}$$

Si observamos los datos con arreglo a este modelo, se comprueba que, efectivamente, la excesiva variabilidad que se observaba dentro de cada tratamiento en el primer ejemplo era debido a que no se había tenido en cuenta las diferentes razas. Teniendo, ahora, en cuenta las razas se observa que la variabilidad dentro de cada casilla es pequeña en referencia a la variabilidad de los tratamientos y de las razas, por lo que, para unos mismos datos, si se analizan con una sola fuente de variación no se observan diferencias entre los diferentes niveles de ese factor, mientras que si se analizan teniendo en cuenta otra fuente de variación, si se observan diferencias entre los niveles del factor en los que antes no se observó diferencias. Por lo que ahora si se podría afirmar con cierta seguridad que el tratamiento de mayor efecto,  $T_2$ , es superior que el tratamiento  $T_1$  y este es superior que el de menor efecto, el  $T_3$ . Esto es como consecuencia de que al introducir el segundo factor se ha aumentado la precisión de medida que ha hecho posible una mayor discriminación de diferencias.

Pero si observamos más detenidamente esta tabla y estimamos los efectos de las casillas restándole a la media de cada casilla la media general y los efectos de los niveles de los factores principales en el que se encuentra la casilla en cuestión, se tiene

$$\begin{aligned} \tau\rho_{11} &= 4.5 - 4.8 - 3 + 3 = -0.3 & \tau\rho_{12} &= 1.5 - 1.2 - 3 + 3 = 0.3 \\ \tau\rho_{21} &= 6.5 - 4.8 - 4 + 3 = -0.7 & \tau\rho_{22} &= 1.5 - 1.2 - 4 + 3 = -0.7 \\ \tau\rho_{31} &= 3.5 - 4.8 - 2 + 3 = -0.3 & \tau\rho_{32} &= 0.5 - 1.2 - 2 + 3 = 0.3 \end{aligned}$$

Con lo que se comprueba que, dentro de la primera raza, el  $T_2$  es claramente mejor que el  $T_1$  y que el  $T_3$ , mientras que estos dos últimos tienen el mismo efecto; y dentro de la segunda raza, el  $T_2$  es el que tiene menor efecto mientras que el  $T_1$  y  $T_3$  tienen el mismo efecto y este es superior al efecto de  $T_2$ .

Para contrastar, supóngase que los datos hubieran sido

	$T_1$	$T_2$	$T_3$			
$R_1$	$X_{111}=6$	$X_{211}=5$	$X_{311}=4$	$X_{.1}=29$	$\bar{X}_{.1}=4.8$	$\rho=1.8$
	$X_{112}=4$	$X_{212}=7$	$X_{312}=3$			
$R_2$	$X_{121}=1$	$X_{221}=3$	$X_{321}=1$	$X_{.2}=7$	$\bar{X}_{.2}=1.2$	$\rho=-1.8$
	$X_{122}=1$	$X_{222}=1$	$X_{322}=0$			
	$X_{1..}=12$	$X_{2..}=16$	$X_{3..}=8$	$X_{...}=36$		
	$\bar{X}_{1..}=3$	$\bar{X}_{2..}=4$	$\bar{X}_{3..}=2$	$\bar{X}_{...}=3$		
	$\tau_1=0$	$\tau_2=1$	$\tau_3=-1$			

Como se ve, las sumas, medias y efectos de los tratamientos y razas son los mismos. También se observa una variabilidad dentro de las casillas más pequeña que la variabilidad debida a los efectos de los tratamientos y las razas, por lo que la conclusión intuitiva, con respecto a los tratamientos y a las razas, es la misma que la de la tabla anterior. Pero si observamos más detenidamente esta tabla y calculamos los efectos de las combinaciones de los dos factores (las casillas), se tiene

$$\begin{aligned} \tau\rho_{11} &= 5 - 4.8 - 3 + 3 = 0.2 & \tau\rho_{12} &= 1 - 1.2 - 3 + 3 = -0.2 \\ \tau\rho_{21} &= 6 - 4.8 - 4 + 3 = 0.2 & \tau\rho_{22} &= 2 - 1.2 - 4 + 3 = -0.2 \\ \tau\rho_{31} &= 3.5 - 4.8 - 2 + 3 = -0.3 & \tau\rho_{32} &= 0.5 - 1.2 - 2 + 3 = 0.3 \end{aligned}$$

Se comprueba que, a diferencia del anterior ejemplo, el  $T_2$  y el  $T_1$  tienen los mismos efectos dentro de cada una de las dos razas, siendo estos efectos superiores al efecto de  $T_3$  en la primera raza e inferiores en la segunda raza.

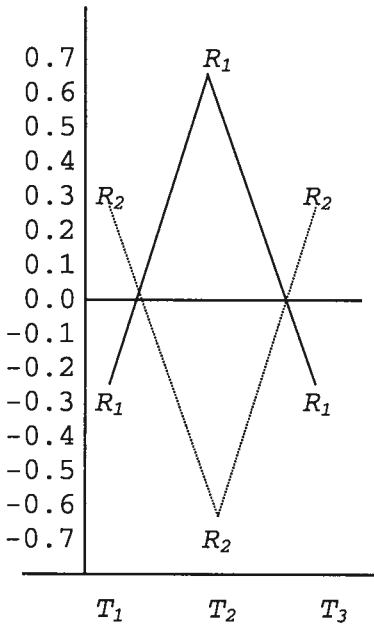
Por lo que no se puede analizar este último ejemplo y el anterior con el mismo modelo lineal, ya que hay que contemplar otra posible fuente de variación (además de *tratamiento* y *raza*) como es la que podría causar que la diferencia entre el mejor tratamiento y el peor tratamiento en una raza sea mucho mayor que la misma diferencia en la otra raza. E incluso podría ocurrir que el mejor tratamiento en una raza no lo sea en otra raza. Es decir, se pueden presentar fenómenos de *sinergia* o de *interferencia* entre los dos factores como consecuencia de que los efectos de estos no sean independientes, es decir, no sean *aditivos*.

## Interacción.-

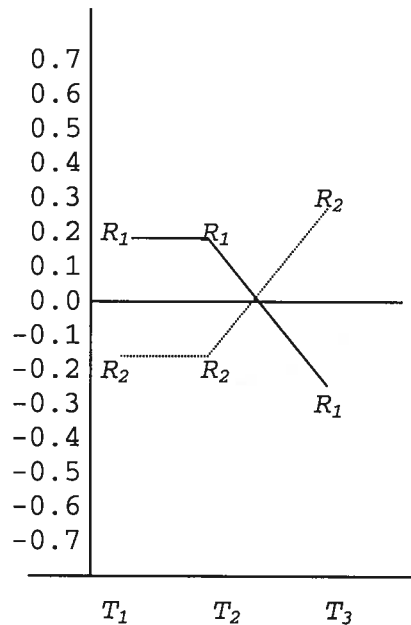
Como consecuencia de esta nueva fuente de variación se introduce un nuevo concepto que va a ser muy útil para la realización correcta de los análisis. Este concepto es el que se conoce con el término estadístico de *interacción*, que como se verá dentro de un momento, no conlleva ninguna complicación irresoluble, pues el modelo se hace aditivo si introducimos esta nueva fuente de variación como un factor más.

La interacción entre dos factores se puede definir como las diferentes tendencias existentes entre los diferentes niveles de un factor dentro de los diferentes niveles del otro factor. Por ejemplo, si se representan los efectos de los tres tratamientos, dentro de cada una de las dos razas, de los dos ejemplos anteriores, se tienen las siguientes figuras

1º ejemplo



2º ejemplo



Como se ve, las líneas que unieran las  $R_1$  se cruzarían con las líneas que unieran las  $R_2$ , lo que significa que no siguen la misma tendencia los efectos de los tres tratamientos dentro de cada una de las dos razas.

Por lo tanto, la relación existente entre los datos de este ejemplo que fue expresada, anteriormente, en términos de desviaciones de la siguiente manera

$$(X_{ijk} - \bar{X}_{...}) = (\bar{X}_{i.} - \bar{X}_{...}) + (\bar{X}_{.j} - \bar{X}_{...}) + (X_{ijk} - \bar{X}_{ij.})$$

no es completa pues no contempla la desviación debida a la posible *interacción* entre los factores.



Esta desviación debida a la interacción es la misma que ya se empleó en el ejemplo anterior (una sola medida por casilla) como error, pues en dicho ejemplo no se podía diferenciar la variabilidad individual de la variabilidad entre casillas al haber una sola medida por casilla, por lo que se utilizó esta desviación, la debida a la interacción, como error, por lo que es muy importante que, en ejemplos con una sola medida por casilla, se esté seguro del supuesto de aditividad, si bien ya se ha visto, en el capítulo anterior, como realizar la *Prueba de no aditividad*.

Esta nueva desviación, por tanto, no es sino la desviación de la media de cada casilla con respecto a la media general menos la desviación del tratamiento y menos la desviación de la raza a la que pertenece la casilla. Esto se puede expresar así

$$TR_{ij} = (X_{ijk} - \bar{X}_{...}) - (\bar{X}_{i.} - \bar{X}_{...}) - (\bar{X}_{.j} - \bar{X}_{...}) = X_{ijk} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{...}$$

es decir, la desviación de cada casilla es el valor medio de dicha casilla menos las medias de los dos grupos a los que pertenece más la media general. Por lo que la desviación total queda de la siguiente manera

$$(X_{ijk} - \bar{X}_{...}) = (\bar{X}_{i.} - \bar{X}_{...}) + (\bar{X}_{.j} - \bar{X}_{...}) + (X_{ijk} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{...}) + (X_{ijk} - \bar{X}_{ij.})$$

Pasando la media general del termino de la derecha al de la izquierda, se obtiene

$$X_{ijk} = \bar{X}_{...} + (\bar{X}_{i.} - \bar{X}_{...}) + (\bar{X}_{.j} - \bar{X}_{...}) + (X_{ijk} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{...}) + (X_{ijk} - \bar{X}_{ij.})$$

Lo que quiere decir que la medida de una unidad experimental es igual a la suma de cinco sumandos (en el ejemplo de una sola medida por casilla eran cuatro sumandos), el primero es la estima de la *media general*,  $\mu$ , el segundo es la estima del efecto del *tratamiento*,  $\tau_i$ , el tercero es la estima del efecto de la *raza*,  $\rho_j$ , el cuarto es la estima de la *interacción*,  $TR_{ij}$ , y el quinto es la estima del *error experimental*,  $\varepsilon_{ij}$ . Por lo que en este caso, el modelo lineal es

$$X_{ijk} = \bar{X}_{...} + (\bar{X}_{i.} - \bar{X}_{...}) + (\bar{X}_{.j} - \bar{X}_{...}) + (\bar{X}_{ij.} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{...}) + (X_{ij.}) - \bar{X}_{ij.}$$

$$X_{ijk} = \mu + T_i + R_j + TR_{ij} + \varepsilon_{ijk}$$

Ahora, en cada individuo o unidad experimental (dato) influye, además de la media general, del efecto del *tratamiento* y del efecto de la *raza*, también influye el posible efecto de la interacción, que en el caso de no tenerla en cuenta en el modelo y ser su variabilidad significativa, estará inflando el efecto correspondiente a los efectos incontrolados ( $e$ ). Es decir, si denominados  $e''$  al error de este modelo,  $e'$  al error del modelo anterior y  $e$  al error del primer modelo, se tiene

$$X_{ij} = \mu + T_i + e_{ij}$$

$$X_{ij} = \mu + T_i + R_j + e'_{ij}$$

$$X_{ijk} = \mu + T_i + R_j + TR_{ij} + e''_{ijk}$$

se tiene que

$$e''_{ij} = e'_{ij} - TR_{ij}$$

ó

$$e''_{ijk} = e_{ij} - R_j - TR_{ij}$$

por lo que el nuevo error será más pequeño cuanto más grandes sean las variabilidades debidas a la segunda fuente de variación introducida en el modelo y a la interacción, por tanto se tendrá mayor precisión de medida.

### **Factores de efectos fijos y factores de efectos aleatorios.-**

Una vez introducido el concepto de interacción y hallado el modelo correcto, hay que introducir otros dos conceptos que nos llevaran a diferenciar tres tipos de análisis para ejemplos de dos factores con repetición (y más adelante para otros tipos de ejemplos). Para ello se van a utilizar los mismos datos que se vienen utilizando desde el primer ejemplo, pero con tres diferentes enunciados.

Enunciado 1: Será el mismo del utilizado en el último ejemplo. Recuérdese que el primer factor se trataba de tres *tratamientos* que podían ser tres dietas o tres antibióticos o tres tóxicos o tres métodos de purificación, etc., es decir, un factor con tres únicos niveles del cual queremos saber si existe diferencias entre dichos niveles y en caso positivo, saber cuál o cuales es el mejor. Con respecto al segundo factor, *raza*, ocurre otro tanto, son dos razas y queremos saber si son diferentes y en caso positivo saber cual es la mejor.

Por lo tanto los efectos de los tratamientos son *fijos*. No hay otros efectos pues son los únicos tratamientos que interesan. Lo mismo para las dos razas.

Como en el modelo que explica los datos de este ejemplo, los dos factores son de efectos fijos, se dice que es un *Modelo Fijo* o *Modelo I*.

Enunciado 2: Supóngase se nos hubiera informado que la experiencia se realizó con animales de dos rebaños diferentes, de manera que el primer factor se trata de tres *tratamientos* del cual queremos saber si existe diferencias entre dichos tratamientos y en caso positivo, saber cuál o cuales es el mejor. Con respecto al segundo factor, *rebaño*, no ocurre lo mismo sino que este segundo factor lo ha introducido el experimentador debida a criterios ambientales o de clasificación. El experimentador no tiene interés en saber si existen diferencias entre los niveles de este segundo factor, de hecho se introduce esta segunda fuente de variación porque se sabe, con cierta certeza, que influye en la variabilidad de la variable que se está midiendo, es decir, se sabe que unos rebaños son diferentes a otros. Como en el

enunciado anterior, el que un animal pertenezca a un rebaño o a otro no es aleatorio mientras que los tratamientos han sido aplicados aleatoriamente a cada animal dentro de cada rebaño.

Aquí hay que introducir un concepto que será de mucha utilidad. Aunque, el que un animal pertenezca a un rebaño o a otro no es aleatorio y si es aleatorio el tratamiento que le corresponde a cada animal, el que se elija un rebaño u otro para realizar la experiencia es *aleatorio*, mientras que los tres tratamientos no son aleatorios, sino que vienen determinados de forma *fija* por el planteamiento del problema. Esto quiere decir, que se ha realizado esta experiencia en estos dos rebaños como se podría haber hecho en cualquiera otros rebaños; y se tiene en cuenta esta fuente de variación porque se sabe que los rebaños (genéricamente, no solo estos dos rebaños concretos) influyen en la variabilidad de la variable que se está midiendo. Mientras que los tratamientos son esos y no otros porque son esos tres los que el experimentador quiere probar, no ha sido una elección aleatoria, sino una imposición del problema planteado.

Lo mismo que se ha dicho de los rebaños se puede decir de las plantas de los hospitales o de los barrios o de las ciudades o de las altitudes, *etc.*

Los rebaños son de efectos aleatorios porque al ser una muestra de los muchos otros rebaños que se podrían haber muestreado, estos efectos hubieran sido diferentes si se hubieran muestreado otros rebaños.

Como en el modelo que explica los datos de este ejemplo, uno de los factores es de efectos fijos y el otro de efectos aleatorios, se dice que es un *Modelo mixto*.

Enunciado 3: En la versión 1 y 2 los tres *Tratamientos* eran fijos, por ejemplo, tres dietas o tres antibióticos o tres tóxicos, *etc.*, es decir, un factor con tres únicos niveles del cual queremos saber si existe diferencias entre ellos y cuál es el mejor. El segundo factor del primer enunciado también era fijo, eran dos razas y se quería saber cuál era la mejor. Pero el segundo factor del segundo enunciado era aleatorio, los dos niveles muestreados es una muestra aleatoria de los muchos niveles (rebaños) existentes de dicho factor. Ahora se va a suponer que los tres niveles del primer factor también van a ser aleatorios, es decir, una muestra aleatoria de los muchos niveles existentes. Por ejemplo, supóngase que se está estudiando la resistencia o capacidad de asimilación del ganado vacuno lechero, de cierto tóxico natural que se produce espontáneamente en el almacenamiento de cierto pienso, para ello se toman cuatro *mediohermanos* hijos de cada uno de tres *machos* (llamémoslo factor *toro* para continuar simbolizándolo con *T*) repartidos dos a dos en dos ganaderías y se le analiza la concentración de dicho tóxico en el hígado. Por tanto, en este caso los tres niveles del primer factor son las tres familias, que es una muestra aleatoria de las muchas familias que se podrían haber utilizado. Siendo los rebaños, como en el segundo enunciado, una muestra aleatoria de los muchos rebaños que se podrían haber utilizado.

Como en el modelo que explica los datos de este ejemplo, los dos factores son de efectos aleatorios, se dice que es un *Modelo aleatorio* o *Modelo II*.

Resumiendo, si los niveles de un factor son la población del total de niveles, se tiene un factor de efectos fijos; si los niveles del factor es una muestra aleatoria de la población de niveles se tiene un factor de efectos aleatorios. Si en un modelo, todos los factores son de efectos fijos, el modelo es fijo o Modelo I; si todos los factores son de efectos aleatorios, el modelo es aleatorio o Modelo II; si se tiene un modelo con unos factores de efectos fijos y otros de efectos aleatorios, se tiene un modelo mixto.

El que se trate de uno u otro tipo de factor depende del problema en sí y del interés del investigador. Puede ser que se esté interesado en estudiar unos tratamientos determinados, por ejemplo tres dosis de abono, cuatro dosis de insecticida, tres temperaturas óptimas de cultivo, tres antibióticos, etc. En estos casos, el factor tratamiento se considera de efectos fijos y el modelo se denomina Modelo I. Pero pudiera ocurrir que nuestro objetivo sea comprobar si un tipo de tratamiento en general tiene influencia sobre la variable, sin estar interesados en ningún nivel especial del factor; por ejemplo, se quiere comprobar si la dieta influye, o si influyen las temperaturas, o si influye el analista en la variable, etc.; los niveles de estos tratamientos a contrastar se tomarán al azar y entonces este factor se dice que es de efectos aleatorios o al azar y el modelo se denomina Modelo II.

De lo dicho extraemos dos conclusiones

- 1) En el caso de un factor fijo las conclusiones del análisis solo se podrán aplicar a los niveles concretos que se han estudiado y nunca a otros posibles niveles.
- 2) Por el contrario, si el factor se consideró aleatorio porque los niveles incluidos fueron elegidos al azar, las conclusiones se aplicarán al factor en sí, independientemente de dichos niveles.

El experimentador, al diseñar la experiencia tiene que saber lo que quiere y por tanto debe conocer o definir si es fijo o aleatorio cada factor considerado. A veces la distinción entre uno y otro no es tan sencilla, pero en general se puede decir que cuando hay un factor dudoso es porque no se está muy interesado en él y su inclusión en el modelo viene impuesta por la experimentación.

Veamos un ejemplo en el que el factor puede ser considerado de efectos fijos o de efectos aleatorios, según el interés de aplicación de los resultados. En el enunciado 2, se ha propuesto un ejemplo con tres tratamientos y dos rebaños, se ha dicho que los rebaños son de efectos aleatorios pues han sido elegidos aleatoriamente de la multitud de rebaños en los que se podían haber realizado la experiencia. Esto es cierto, pero podría haber ocurrido que la experiencia hubiera sido la de probar los tres tratamientos en dos rebaños de nuestra propiedad o de la persona que nos ha encargado el estudio; es decir, estamos interesados concretamente en esos dos rebaños y las conclusiones son para ellos dos, no para los rebaños de otra personas.

Lo mismo que se ha dicho de los rebaños podría decirse de las plantas del hospital o de los hospitales. Puede ser que, para probar los tres tratamientos, se haya elegido esos dos hospitales o dos plantas porque el problema esté planteado en esas dos plantas o dos hospitales, y las conclusiones son para ellos no para el resto de los hospitales.

## Modelo de efectos fijos.-

Corresponde al del enunciado 1 en el que ambos factores son de efectos fijos. Se va a utilizar los mismos datos del ejemplo primero. Recuérdese que el primer factor se trataba de tres *tratamientos* que podían ser tres dietas o tres antibióticos o tres tóxicos o tres métodos de purificación, *etc.*, es decir, un factor con tres únicos niveles del cual queremos saber si existe diferencias entre dichos niveles y en caso positivo, saber cuál o cuales es el mejor. Con respecto al segundo factor, *raza*, ocurre otro tanto, son dos razas y queremos saber si son diferentes y en caso positivo saber cual es la mejor. Por lo tanto los efectos de los tratamientos son *fijos*. No hay otros efectos pues son los únicos tratamientos que interesan. Lo mismo para las dos razas. Como en el modelo que explica los datos de este ejemplo, los dos factores son de efectos fijos, se dice que es un *Modelo Fijo* o *Modelo I*.

Como ambos factores son de efectos fijos, se tiene que

$$\begin{aligned}\sum_i T_i &= 0 \\ \sum_j R_j &= 0 \\ \sum_i TR_{ij} &= 0 \\ \sum_j TR_{ij} &= 0 \\ e_{ijk} &\sim N(0, \sigma^2)\end{aligned}$$

por lo que la única varianza que se tiene es la de la población que, como en todos los demás modelos, es estimada por el cuadrado medio del error. Por lo que al estimar la variabilidad de las tres fuentes de variación, se estará haciendo a través de la variabilidad del error, esto hace que estas variabilidades incluyan, además de la debida a dichos factores, la variabilidad del error, con la que se va a comparar para saber si los tratamientos, las razas y la interacción son diferentes o no.

Efectivamente, si vemos la variabilidad que hay dentro de cada casilla (o variabilidad del error) se observa que es mucho menor que la variabilidad debida a los tratamientos y que la debida a las razas, por lo que, ahora, se puede afirmar con cierta certeza que, al menos, el segundo tratamiento (de mayor efecto) es mejor que el tercero (de menor efecto); mientras que cuando se observaban estos datos con solo los tratamientos (sin las razas) no se estaba seguro de poder afirmar esto.

Mientras que para la interacción, es decir, la variabilidad de las casillas después de haberle quitado la de los factores principales (ver la desviación debida a la interacción en la formulación de este modelo), no se puede afirmar lo mismo con la misma seguridad, pues aunque las desviaciones que se observan entre casillas son mayores que las que se observan dentro de las casillas, no son los suficientes como para afirmar, intuitivamente, que existen diferencias (más adelante se hará la prueba).

## Modelo de efectos aleatorios.-

Corresponde al del enunciado 3 en el que ambos factores son de efectos aleatorios. Se va a utilizar los mismos datos del ejemplo primero y, por lo tanto, del ejemplo anterior. Recuérdese que en este caso se esta estudiando la resistencia o capacidad de asimilación del ganado vacuno lechero, de cierto tóxico natural que se produce espontáneamente en el almacenamiento de cierto pienso, para ello se toman cuatro *mediohermanos* hijos de cada uno de tres *machos* repartidos dos a dos en dos ganaderías y se le analiza la concentración de dicho tóxico en el hígado. Por tanto, en este caso los tres tratamientos o tres niveles del primer factor son las tres familias, que es una muestra aleatoria de las muchas familias que se podrían haber utilizado. Siendo los rebaños, como en el segundo enunciado, una muestra aleatoria de los muchos rebaños que se podrían haber utilizado. Como en el modelo que explica los datos de este ejemplo, los dos factores son de efectos aleatorios, se dice que es un *Modelo aleatorio* o *Modelo II*.

Como ambos factores son de efectos aleatorios, se tiene que

$$T_i \sim N(0, \sigma_T^2)$$

$$R_j \sim N(0, \sigma_R^2)$$

$$TR_{ij} \sim N(0, \sigma_{TR}^2)$$

$$e_{ijk} \sim N(0, \sigma^2)$$

por lo que, en este caso, se tiene cuatro varianzas: la del factor  $T$ , la del factor  $R$ , la de la interacción y la del error, que es la de la población. Por lo que al estimar la variabilidad de los tratamientos y de los rebaños se hará a través de la variabilidad de la interacción y la del error, por lo que incluirá, además de la debida a dichos factores, la variabilidad de la interacción y la variabilidad del error; mientras que la variabilidad de la interacción incluirá su variabilidad y la del error.

Por tanto, en este caso, hay que comparar la variabilidad de los tratamientos y de los rebaños con la de la interacción y la de esta con la del error. Si se hace así se comprueba que la variabilidad de los tratamientos no es mucho mayor que la de interacción, por lo que ya no estaríamos tan seguros de afirmar diferencias entre tratamientos como lo estábamos en el modelo de efectos fijos. Para rebaños si seguimos estado seguros porque la variabilidad de estos es mucho mayor. La interacción, como en el modelo anterior, no presenta una variabilidad manifiestamente mayor que la de los interiores de las casillas.

Se vuelve a repetir que se está tratando de hacer un análisis intuitivo, el análisis estadístico se hará más adelante.

## Modelo mixto.-

Corresponde al del enunciado 2 en el que un factor es de efectos fijos y el otro de efectos aleatorios. Se va a utilizar los mismos datos del ejemplo primero y, por lo tanto, los mismos de los dos últimos ejemplos. Recuérdese que el primer factor se trataba de tres *tratamientos* que podían ser tres dietas o tres antibióticos o tres tóxicos o tres métodos de purificación, *etc.*, es decir, un factor con tres únicos niveles del cual queremos saber si existe diferencias entre dichos niveles y en caso positivo, saber cuál o cuales es el mejor. Con respecto al segundo factor, *rebaño*, no ocurre lo mismo sino que este segundo factor lo ha introducido el experimentador obligado por criterios ambientales o de clasificación. El experimentador no tiene interés en saber si existen diferencias entre los niveles de este segundo factor, de hecho se introduce esta segunda fuente de variación porque se sabe, con cierta certeza, que influye en la variabilidad de la variable que se está midiendo, es decir, se sabe que unos rebaños son diferentes a otros. Y en todo caso, si no se sabe si el factor rebaño influye en esta variable, lo que se pretende saber es si es significativo como factor, no si estos dos rebaños concretos son diferentes o no.

Como en el modelo que explica los datos de este ejemplo, un factor es de efectos fijos y el otro es aleatorio, se dice que es un *Modelo mixto*.

Como  $R$  es de efectos aleatorios, se tiene que los efectos del factor  $T$  (fijo) se estiman a través de los niveles del factor  $R$  que es aleatorio, por lo que los efectos de las casillas dentro de los diferentes niveles del factor fijo son una muestra aleatoria de casillas que se ajustarán a una distribución normal de media cero y varianza la de la interacción. Mientras que los efectos del factor  $R$  (aleatorio) se estiman a través de los diferentes niveles del factor  $T$  que es fijo, por lo que los efectos de las casillas dentro de los diferentes niveles del factor aleatorio son fijos, por lo que tienen que valer cero y no contienen varianza ninguna pues son fijos.

$$\sum_i T_i = 0$$

$$R_j \sim N(0, \sigma_R^2)$$

$$TR_{ij} \sim N(0, \sigma_{TR}^2)$$

$$\sum_i TR_{ij} = 0$$

$$\sum_j TR_{ij} \sim N(0, \sigma_{TR}^2)$$

$$e_{ijk} \sim N(0, \sigma^2)$$

por lo que, ahora, se tiene tres varianzas: la del factor aleatorio, rebaño, la de la interacción y la del error que estima la de la población.

Como

$$\sum_i TR_{ij} = 0$$

Indica que son fijos los efectos de las casillas dentro de cada nivel del factor  $R$ , es decir, del factor aleatorio. Por lo que la variabilidad de los rebaños no incluirá la variabilidad de la interacción porque esta es de efectos fijos para dicho factor. Por lo que la variabilidad de rebaño incluye, además de la debida a dicho factor, la debida al error.

Y como

$$\sum_i TR_{ij} \sim N(0, \sigma_{TR}^2)$$

Indica que son aleatorios los efectos de las casillas dentro de cada nivel del factor  $T$ , es decir, del factor fijo. Por lo que la variabilidad de los tratamientos incluirá la variabilidad de la interacción porque esta es de efectos aleatorios para dicho factor. Por lo que la variabilidad de los tratamientos incluirá, además de la debida a dicho factor, la variabilidad de la interacción y la del error.

Por tanto, en este caso, hay que comparar la variabilidad de los tratamientos con la de la interacción, y la de los rebaños con la del error. Si se hace así se comprueba que la variabilidad de los tratamientos no es mucho mayor que la de interacción, por lo que ya no estaríamos tan seguros de afirmar diferencias entre tratamientos como lo estábamos en el modelo de efectos fijos. Para los rebaños si seguimos estado seguros porque la variabilidad es mucho mayor. La interacción, como en los modelos anteriores, no presenta una variabilidad manifiestamente mayor que la de los interiores de las casillas o error.

Se insiste en que se está tratando de hacer un análisis intuitivo, el análisis estadístico se hará ahora.

### **Análisis de varianza Factorial con repetición.-**

Como se ha dicho anteriormente, en el análisis del diseño factorial sin repetición se utiliza como término de error el de la interacción. Si se quiere diferenciar la variabilidad entre casillas de la variabilidad individual, se necesita varias medidas por casilla, puesto que si se tiene una sola medida por casilla la variabilidad individual coincide con la variabilidad entre casillas.

Esto es algo que ocurre con cierta frecuencia, pues como se vio en el epígrafe *Número de medidas por casilla*, cuando se tenga un modelo lineal con dos factores puede ocurrir a menudo que se tienen varias medidas por casilla como consecuencia de que se tienen más de  $t$  unidades experimentales por bloque. En el ejemplo anterior se puede tener más de tres animales por raza para realizar la experiencia; por ejemplo, supóngase que se ha realizado la experiencia en seis animales por raza, por lo que cada tratamiento se repite dos veces en cada raza.

### **Modelo Factorial con repetición.-**

Un diseño factorial con varias medidas por casilla es el que se ajusta al siguiente modelo lineal



$$X_{ijk} = \mu + T_i + R_j + TR_{ij} + e_{ijk}$$

donde  $\mu$  es la media general,  $T_i$  es el efecto del tratamiento  $i$ ,  $R_j$  es el efecto de la raza  $j$  y  $TR_{ij}$  expresa el efecto de la interacción entre el tratamiento  $i$  y la raza  $j$ , y  $e_{ijk}$  es el efecto del individuo  $k$  dentro de la combinación  $ij$ .

### Descomposición de la varianza y de los grados de libertad en un modelo factorial con repetición.-

Los sumandos del modelo anterior vienen estimados de la siguiente manera

$$\mu = \bar{X}_{...}$$

El efecto del tratamiento  $i$  será el valor medio de dicho tratamiento menos lo debido a la media general. Y lo mismo para la  $j$ -ésima raza

$$T_i = \bar{X}_{i..} - \mu = \bar{X}_{i..} - \bar{X}_{...}$$

$$R_j = \bar{X}_{.j.} - \mu = \bar{X}_{.j.} - \bar{X}_{...}$$

La estima del efecto de la  $ij$ -ésima casilla es restándole, además de la media general, el efecto del tratamiento  $i$ , de la raza  $j$ , con lo que se obtiene el efecto puro de la interacción.

$$\begin{aligned} TR_{ij} &= \bar{X}_{ij.} - T_i - R_j - \mu = \bar{X}_{ij.} - (\bar{X}_{i..} - \bar{X}_{...}) - (\bar{X}_{.j.} - \bar{X}_{...}) - \bar{X}_{...} = \\ &= \bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...} \end{aligned}$$

Para la estima del error se sigue un concepto análogo

$$\begin{aligned} \varepsilon_{ijk} &= X_{ijk} - TR_{ij.} - T_i - R_j - \mu = \\ &= X_{ijk} - (\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...}) - (\bar{X}_{i..} - \bar{X}_{...}) - (\bar{X}_{.j.} - \bar{X}_{...}) = \\ &= X_{ijk} - \bar{X}_{ij.} \end{aligned}$$

Así pues la desviación entre una medida cualquiera,  $X_{ijk}$ , y la media general, se puede descomponer de la siguiente forma

$$X_{ijk} - \bar{X}_{...} = (\bar{X}_{i..} - \bar{X}_{...}) + (\bar{X}_{.j.} - \bar{X}_{...}) + (\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...}) + (X_{ijk} - \bar{X}_{ij.})$$

De forma similar a lo realizado en los modelos anteriores (elevando al cuadrado, sumando para  $ijk$ , etc.) se obtiene la siguiente descomposición de la suma de cuadrados total, con los grados de libertad correspondientes

FV	gl	SC
Trata	t-1	$\frac{1}{m} \sum_i X_{i..}^2 - \frac{X_{...}^2}{N}$
Raza	r-1	$\frac{1}{tn} \sum_j X_{.j.}^2 - \frac{X_{...}^2}{N}$
TxR	(t-1)(r-1)	$\frac{1}{n} \sum_{ij} X_{ij}^2 - \frac{1}{m} \sum_i X_{i..}^2 - \frac{1}{tn} \sum_j X_{.j.}^2 + \frac{X_{...}^2}{N}$
Error	N-tr	$\sum_{ijk} X_{ij}^2 - \frac{1}{n} \sum_{ij} X_{ij}^2$
Total	N-1	$\sum_{ijk} X_{ijk}^2 - \frac{X_{...}^2}{N}$

siendo  $N = tm$

En la composición de estas expresiones se ve, como en los modelos anteriores, que hay una lógica, cada sumando elevado al cuadrado está dividido por el número de elementos que han formado esta suma, por ejemplo, si han intervenido  $m$  números se dividirá por  $m$ ; si la suma es de  $tn$  individuos, hay que dividir por  $tn$ ; y si la suma es de  $n$  individuos, hay que dividir por  $n$ , etc. Y lo mismo se puede decir en lo concerniente al concepto de grados de libertad.

### Parámetros estimados en un Modelo Factorial con repetición.-

Cuando se han realizado pruebas de hipótesis para modelos de una vía y modelos factoriales con un solo dato por casilla, las  $F$  de prueba son siempre las mismas, sea cual sea el tipo o tipos de factores que se tengan. Pero en el caso del modelo factorial con repetición, las esperanzas matemáticas de los cuadrados medios de los diferentes factores son diferentes según el tipo de factor.

Por tanto, antes de continuar con los parámetros estimados por los cuadrados medios, repasemos lo que son *factores de efectos fijos* y *factores de efectos aleatorios*.

Ya se ha dicho en epígrafes anteriores que el que se trate de uno u otro tipo de factor depende del problema en sí y del interés del investigador. Puede ser que se esté interesado en estudiar unos tratamientos determinados, por ejemplo tres dosis de abono, cuatro dosis de insecticida, tres temperaturas óptimas de cultivo, tres antibióticos, etc. En estos casos, el factor tratamiento se considera de efectos fijos y el modelo se denomina *Modelo I*. Pero pudiera ocurrir que el objetivo sea comprobar si un tipo de tratamiento, en general, tiene influencia sobre la producción, sin estar interesados en ningún nivel especial del factor; por ejemplo, se quiere comprobar si la dieta influye, o si influyen las temperaturas en la producción o en el crecimiento de los niños, etc.; los niveles de estos tratamientos a contrastar se tomarán al azar y entonces

este factor se dice que es de efectos aleatorios o al azar y el modelo se denomina *Modelo II*.

De lo dicho se extraen dos conclusiones

- 1) En el caso de un factor fijo las conclusiones del análisis solo se podrán aplicar a los niveles concretos que se han estudiado y nunca a otros posibles niveles.
- 2) Por el contrario, si el factor se consideró aleatorio y los niveles incluidos fueron elegidos al azar, las conclusiones se aplicarán al factor en sí, independientemente de dichos niveles.

El experimentador, al diseñar la experiencia tiene que saber lo que quiere y por tanto debe conocer o definir el carácter de fijo o aleatorio para cada factor considerado. A veces la distinción entre uno y otro no es tan sencilla, pero en general se puede decir que cuando hay un factor dudoso es porque no se está muy interesado en él y se esté considerando como consecuencia de imposiciones de la experimentación.

Supongamos, primero, que se trata de un *Modelo fijo* (ver epígrafe) correspondiente al enunciado 1 de los vistos en el epígrafe *Factores de efectos fijos y factores de efectos aleatorios*. En este caso los tres niveles del primer factor (tres tratamientos) son los únicos niveles que existen o que nos interesan y se quiere saber si existen diferencias entre dichos niveles. Y para el segundo factor ocurre otro tanto, los dos niveles (dos razas) son las únicas que existen o que nos interesan y se quiere saber si son iguales o diferentes.

En el modelo fijo, tanto los *tratamientos* como las *razas* se estiman a través de todo el universo de valores de razas y tratamientos, respectivamente, por lo que la única varianza que se tiene es la de la población que, como en todos los demás modelos, es estimada por el error de manera que

$$\begin{aligned}\sum_i T_i &= 0 \\ \sum_j R_j &= 0 \\ \sum_i TR_{ij} &= 0 \\ \sum_j TR_{ij} &= 0 \\ e_{ijk} &\sim N(0, \sigma^2)\end{aligned}$$

Por lo que al estimar la variabilidad de las tres fuentes de variación (*tratamiento, raza e interacción*), se estará haciendo a través de la variabilidad del error, esto hace que estas variabilidades incluyan, además de la debida a dichos factores, la variabilidad del error.

Calculando, por tanto, los valores que estiman las sumas de cuadrados de manera análoga a como se hizo con el modelo de una sola vía, y al dividir las estimas de las *SC* por los *gl* se obtienen las de los *CM*, que son

$$\begin{aligned}
 ECM_{Tratamientos} &= rn \sigma_T^2 + \sigma^2 \\
 ECM_{Razas} &= tn \sigma_R^2 + \sigma^2 \\
 ECM_{TR} &= n \sigma_{TR}^2 + \sigma^2 \\
 ECM_{Error} &= \sigma^2
 \end{aligned}$$

Quedando la tabla resumen de la siguiente manera

FV	gl	Modelo Fijo	
		SC	ECM
Trata	t-1	$\frac{1}{rn} \sum_i X_{i..}^2 - \frac{X_{...}^2}{N}$	$\sigma^2 + rn \sigma_T^2$
Raza	r-1	$\frac{1}{tn} \sum_j X_{.j.}^2 - \frac{X_{...}^2}{N}$	$\sigma^2 + tn \sigma_R^2$
TxR	(t-1)(r-1)	$\frac{1}{n} \sum_{ij} X_{ij}^2 - \frac{1}{rn} \sum_i X_{i..}^2 - \frac{1}{tn} \sum_j X_{.j.}^2 + \frac{X_{...}^2}{N}$	$\sigma^2 + n \sigma_{TR}^2$
Error	N-tr	$\sum_{ijk} X_{ijk}^2 - \frac{1}{n} \sum_{ij} X_{ij.}^2$	$\sigma^2$
Total	N-1	$\sum_{ijk} X_{ijk}^2 - \frac{X_{...}^2}{N}$	

Supóngase, ahora, que se tiene un *Modelo Aleatorio* (ver epígrafe) correspondiente al enunciado 3 de los vistos en el epígrafe *Factores de efectos fijos y factores de efectos aleatorios*. En este caso los tres niveles del primer factor o los tres tratamientos (tres *familias*) son una muestra aleatoria de los muchos niveles de este factor, y los dos niveles segundo factor (dos *rebaños*) son, también, una muestra aleatoria de los muchos niveles existentes. Por lo que se tienen cuatro varianzas, estas son

$$\begin{aligned}
 T_i &\sim N(0, \sigma_T^2) \\
 R_j &\sim N(0, \sigma_R^2) \\
 TR_{ij} &\sim N(0, \sigma_{TR}^2) \\
 e_{ijk} &\sim N(0, \sigma^2)
 \end{aligned}$$

En el modelo aleatorio, por tanto, la estima de los efectos de un factor se realiza a través de una muestra aleatoria del universo de niveles del otro factor, y viceversa.

Por lo que al estimar la variabilidad de los factores principales (*familia, rebaño*), se estará haciendo a través de la variabilidad de las casillas (interacción) y de la variabilidad individual (error), esto hace que las variabilidades de familia y de rebaño, incluyan, además de la debida a dichos factores, la variabilidad de la interacción y del

error. Mientras que la estima de los efectos de las casillas se hará, como antes, a través de la variabilidad individual, por lo que incluirá, además de la varianza de la interacción, la varianza del error.

Calculando, por tanto, los valores que estiman las sumas de cuadrados de manera análoga a como se hizo con el modelo de una sola vía, y al dividir las estimas de las *SC* por los *gl* se obtienen las de los *CM*, que son

$$\begin{aligned}
 ECM_{Familias} &= n\sigma_T^2 + n\sigma_{TR}^2 + \sigma^2 \\
 ECM_{Rebaños} &= tn\sigma_R^2 + n\sigma_{TR}^2 + \sigma^2 \\
 ECM_{FR} &= n\sigma_{TR}^2 + \sigma^2 \\
 ECM_{Error} &= \sigma^2
 \end{aligned}$$

Quedando la tabla resumen de la siguiente manera

<b>Modelo Aleatorio</b>			
<i>FV</i>	<i>gl</i>	<i>SC</i>	<i>ECM</i>
<i>Trata</i>	<i>t-1</i>	$\frac{1}{m} \sum_i X_{i..}^2 - \frac{X_{...}^2}{N}$	$\sigma^2 + n\sigma_{TR}^2 + m\sigma_T^2$
<i>Raza</i>	<i>r-1</i>	$\frac{1}{tn} \sum_j X_{.j.}^2 - \frac{X_{...}^2}{N}$	$\sigma^2 + n\sigma_{TR}^2 + tn\sigma_R^2$
<i>TxR</i>	<i>(t-1)(r-1)</i>	$\frac{1}{n} \sum_{ij} X_{ij}^2 - \frac{1}{m} \sum_i X_{i..}^2 - \frac{1}{tn} \sum_j X_{.j.}^2 + \frac{X_{...}^2}{N}$	$\sigma^2 + n\sigma_{TR}^2$
<i>Error</i>	<i>N-tr</i>	$\sum_{ijk} X_{ij}^2 - \frac{1}{n} \sum_{ij} X_{ij}^2$	$\sigma^2$
<i>Total</i>	<i>N-1</i>	$\sum_{ijk} X_{ijk}^2 - \frac{X_{...}^2}{N}$	

Supóngase, ahora, que se tiene un *Modelo Mixto* (ver epígrafe) correspondiente al enunciado 2 de los vistos en el epígrafe *Factores de efectos fijos y factores de efectos aleatorios*. En este caso los tres niveles del primer factor (tres *tratamientos*) son el universo de niveles de este factor, no hay más tratamientos o no nos interesan otros, si los hubiera, y los dos niveles segundo factor (dos *rebaños*) son una muestra aleatoria de los muchos niveles existentes. Por lo que se tienen cuatro varianzas, estas son

$$\begin{aligned} \sum_i T_i &= 0 \\ R_j &\sim N(0, \sigma_R^2) \\ TR_{ij} &\sim N(0, \sigma_{TR}^2) \\ \sum_i TR_{ij} &= 0 \\ \sum_j TR_{ij} &\sim N(0, \sigma_{TR}^2) \\ e_{ijk} &\sim N(0, \sigma^2) \end{aligned}$$

En el modelo aleatorio, por tanto, la estima de los efectos del factor fijo se realiza a través de una muestra aleatoria del universo de niveles del factor aleatorio, y la estima de los efectos del factor aleatorio se realiza a través del universo de niveles del factor fijo. Por lo que al estimar la variabilidad del factor fijo (*tratamientos*), se estará haciendo a través de la variabilidad de las casillas (interacción) y de la variabilidad individual (error), esto hace que la variabilidad debida a los tratamientos incluya, además de la debida a dicho factor, la variabilidad de la interacción y del error. Mientras que al estimar la variabilidad del factor aleatorio (*rebaño*), se estará haciendo a través de la variabilidad individual, pues al ser

$$\sum_i TR_{ij} = 0$$

se anulan los efectos de las casillas (interacción), esto hace que la variabilidad debida a los rebaños incluya, además de la variabilidad debida a dicho factor, la variabilidad del error. La estima de los efectos de las casillas se hará, como siempre, a través de la variabilidad individual, por lo que incluirá, además de la varianza de la interacción, la varianza del error.

Calculando, por tanto, los valores que estiman las sumas de cuadrados de manera análoga a como se hizo con el modelo de una sola vía, y al dividir las estimas de las *SC* por los *gl* se obtienen las de los *CM*, que son

$$\begin{aligned} ECM_{Tratamientos} &= rn \sigma_T^2 + n \sigma_{TR}^2 + \sigma^2 \\ ECM_{Rebaño} &= tn \sigma_R^2 + \sigma^2 \\ ECM_{TR} &= n \sigma_{TR}^2 + \sigma^2 \\ ECM_{Error} &= \sigma^2 \end{aligned}$$

Quedando la tabla resumen de la siguiente manera

<b>Modelo Mixto</b>			
<i>FV</i>	<i>gl</i>	<i>SC</i>	<i>ECM</i>
<i>Trata</i>	<i>t-1</i>	$\frac{1}{m} \sum_i X_{i..}^2 - \frac{X_{...}^2}{N}$	$\sigma^2 + n\sigma_{TR}^2 + m\sigma_T^2$
<i>Raza</i>	<i>r-1</i>	$\frac{1}{tn} \sum_j X_{.j.}^2 - \frac{X_{...}^2}{N}$	$\sigma^2 + tn\sigma_R^2$
<i>TxR</i>	$(t-1)(r-1)$	$\frac{1}{n} \sum_{ij} X_{ij}^2 - \frac{1}{m} \sum_i X_{i..}^2 - \frac{1}{tn} \sum_j X_{.j.}^2 + \frac{X_{...}^2}{N}$	$\sigma^2 + n\sigma_{TR}^2$
<i>Error</i>	<i>N-tr</i>	$\sum_{ijk} X_{ijk}^2 - \frac{1}{n} \sum_{ij} X_{ij.}^2$	$\sigma^2$
<i>Total</i>	<i>N-1</i>	$\sum_{ijk} X_{ijk}^2 - \frac{X_{...}^2}{N}$	

Insistamos en que esta aparente paradoja de que, en el modelo mixto, la interacción aparece en el efecto fijo, mientras que en el modelo fijo no aparece y sí en el modelo aleatorio, se debe a que el factor fijo se estima a través de datos aleatorios y por tanto como hemos visto antes no se anula el efecto de la interacción; mientras que el factor aleatorio se estima a través de todos los niveles del otro factor y en ese caso sí se anula el efecto de la interacción.

### Pruebas de hipótesis en un modelo factorial con repetición.-

Como en el modelo anterior dos posibles pruebas de hipótesis son las de contrastar si los efectos debidos a los diferentes niveles de los dos factores son o no significativos y, se puede añadir en este caso, la prueba para contrastar si existe interacción entre ambos factores, por tanto las pruebas de hipótesis pueden ser

Para el primer factor

$$H_0 : \sigma_T^2 = 0$$

$$H_1 : \sigma_T^2 > 0$$

Para el segundo factor

$$H_0 : \sigma_R^2 = 0$$

$$H_1 : \sigma_R^2 > 0$$

Y para la interacción

$$H_0 : \sigma_{TR}^2 = 0$$

$$H_1 : \sigma_{TR}^2 > 0$$

Para realizar la prueba  $F$ , que determine la posible significación de un factor o fuente de variación (que la varianza producida por el factor es mayor de cero) se toma como numerador el  $CM$  de dicho factor y como denominador el  $CM$  de la fuente de variación que en su  $ECM$  presenta las mismas componentes de la varianza del numerador, afectadas de iguales coeficientes, menos la componente que se está contrastando si es igual a cero.

El  $CM$  que se usa como denominador para la prueba  $F$  se suele denominar *término de error* para el factor que se está probando; que no tiene por qué ser el término llamado *error* que es realmente el que mide la variabilidad residual.

Por tanto, las posibles pruebas  $F$  que se pueden realizar en modelos factoriales con repetición, son diferentes según el tipo de factores que se tengan, pues según el tipo de factor, sus cuadrados medios estimarán diferentes parámetros y por lo tanto, tendrán como término de error diferentes cuadrados medios.

La generalización para la elaboración de los parámetros estimados por los cuadrados medios de los factores de un factorial (o de cualquier otro modelo) con objeto, entre otras finalidades, de saber cuales son las pruebas de significación para las diferentes fuentes de variación, se verá más adelante, en el epígrafe *Cálculo de las estimas de los CM por el método rápido*.

Empecemos por el Modelo Fijo.

En el caso de la prueba del primer factor, si  $\sigma_T^2 = 0$

$$ECM_{Tratamiento} = m \sigma_T^2 + \sigma^2 = m \cdot 0 + \sigma^2 = \sigma^2 = ECM_{Error}$$

Es decir, el cuadrado medio de los tratamientos estima lo mismo que el cuadrado medio del error, luego, si se dividen ambos cuadrados medios, se espera un valor de  $F$  no significativo, esto es, estadísticamente igual a uno.

Y si  $\sigma_T^2 > 0$

$$ECM_{Tratamiento} = m \sigma_T^2 + \sigma^2 > \sigma^2 = ECM_{Error}$$

Es decir, el cuadrado medio de los tratamientos estima una componente más (la debida a los tratamientos) que el cuadrado medio del error, luego, si se dividen ambos cuadrados medios, se espera un valor de  $F$  significativo, esto es, estadísticamente superior a uno.

Por lo que el estadístico de prueba para el primer factor será



$$F_o = \frac{CM_{Tratamiento}}{CM_{Error}}$$

En el caso de la prueba del segundo factor, si  $\sigma^2_R = 0$

$$ECM_{Raza} = tn \sigma_R^2 + \sigma^2 = tn 0^2 + \sigma^2 = \sigma^2 = ECM_{Error}$$

Es decir, el cuadrado medio de las razas estima lo mismo que el cuadrado medio del error, luego, si se dividen ambos cuadrados medios, se espera un valor de  $F$  no significativo, esto es, estadísticamente igual a uno.

Y si  $\sigma^2_R > 0$

$$ECM_{Raza} = tn \sigma_R^2 + \sigma^2 > \sigma^2 = ECM_{Error}$$

Es decir, el cuadrado medio de las razas estima una componente más (la debida a las razas) que el cuadrado medio del error, luego, si se dividen ambos cuadrados medios, se espera un valor de  $F$  significativo, esto es, estadísticamente superior a uno.

Por lo que el estadístico de prueba para el segundo factor será

$$F_o = \frac{CM_{Raza}}{CM_{Error}}$$

En el caso de la prueba de la tercera fuente de variación, esta es, la de la interacción, si  $\sigma^2_{TR} = 0$

$$ECM_{TR} = n \sigma_{TR}^2 + \sigma^2 = n 0^2 + \sigma^2 = ECM_{Error}$$

Es decir, el cuadrado medio de la interacción estima lo mismo que el cuadrado medio del error, luego, si se dividen ambos cuadrados medios, se espera un valor de  $F$  no significativo, esto es, estadísticamente igual a uno.

Y si  $\sigma^2_{TR} > 0$

$$ECM_{TR} = n \sigma_{TR}^2 + \sigma^2 > \sigma^2 = ECM_{Error}$$

Es decir, el cuadrado medio de la interacción estima una componente más (la debida a la interacción) que el cuadrado medio del error, luego, si se dividen ambos cuadrados medios, se espera un valor de  $F$  significativo, esto es, estadísticamente superior a uno.

Por lo que el estadístico de prueba para la tercera fuente de variación o interacción es

$$F_o = \frac{CM_{TR}}{CM_{Error}}$$

Quedando el análisis de esta manera

### Modelo Fijo

FV	gl	SC	CM	ECM	$F_o$
Trata	t-1	$\frac{1}{m} \sum_i X_{i..}^2 - \frac{X_{...}^2}{N}$	$\frac{SC_{Tra}}{t-1}$	$\sigma^2 + m \sigma_T^2$	$\frac{CM_{Tra}}{CM_{Err}}$
Raza	r-1	$\frac{1}{tn} \sum_j X_{.j.}^2 - \frac{X_{...}^2}{N}$	$\frac{SC_{Raz}}{r-1}$	$\sigma^2 + tn \sigma_R^2$	$\frac{CM_{Raz}}{CM_{Err}}$
TxR	(t-1)(r-1)	$\frac{1}{n} \sum_{ij} X_{ij}^2 - \frac{1}{m} \sum_i X_{i..}^2 - \frac{1}{tn} \sum_j X_{.j.}^2 + \frac{X_{...}^2}{N}$	$\frac{SC_{TR}}{(t-1)(r-1)}$	$\sigma^2 + n \sigma_{TR}^2$	$\frac{CM_{TR}}{CM_{Err}}$
Error	N-tr	$\sum_{ijk} X_{ijk}^2 - \frac{1}{n} \sum_{ij} X_{ij.}^2$	$\frac{SC_{Err}}{N-tr}$	$\sigma^2$	
Total	N-1	$\sum_{ijk} X_{ijk}^2 - \frac{X_{...}^2}{N}$			

Veamos ahora el Modelo Aleatorio.

Recordemos que el enunciado que nos ha servido de ejemplo, el primer factor eran familias y el segundo factor eran rebaños, pero sigamos simbolizando el primer factor como T.

En el caso de la prueba del primer factor, si  $\sigma_T^2 = 0$

$$ECM_{Trata} = m\sigma_T^2 + n\sigma_{TR}^2 + \sigma^2 = m0^2 + n\sigma_{TR}^2 + \sigma^2 = n\sigma_{TR}^2 + \sigma^2 = ECM_{TR}$$

Es decir, el cuadrado medio del primer factor estima lo mismo que el cuadrado medio de la interacción, luego, si se dividen ambos cuadrados medios, se espera un valor de F no significativo, esto es, estadísticamente igual a uno.

Y si  $\sigma_T^2 > 0$

$$ECM_{Trata} = m\sigma_T^2 + n\sigma_{TR}^2 + \sigma^2 > n\sigma_{TR}^2 + \sigma^2 = ECM_{TR}$$

Es decir, el cuadrado medio de los tratamientos estima una componente más (la debida a los tratamientos) que el cuadrado medio de la interacción, luego, si se dividen ambos cuadrados medios, se espera un valor de F significativo, esto es, estadísticamente superior a uno.

Por lo que el estadístico de prueba para el primer factor será

$$F_o = \frac{CM_{Trata}}{CM_{TR}}$$

En el caso de la prueba del segundo factor, si  $\sigma_R^2 = 0$

$$ECM_{Rebaño} = tn\sigma_R^2 + n\sigma_{TR}^2 + \sigma^2 = tn0^2 + n\sigma_{TR}^2 + \sigma^2 = n\sigma_{TR}^2 + \sigma^2 = ECM_{TR}$$

Es decir, el cuadrado medio de los rebaños estima lo mismo que el cuadrado medio de la interacción, luego, si se dividen ambos cuadrados medios, se espera un valor de  $F$  no significativo, esto es, estadísticamente igual a uno.

Y si  $\sigma_R^2 > 0$

$$ECM_{Rebaño} = tn\sigma_R^2 + n\sigma_{TR}^2 + \sigma^2 > n\sigma_{TR}^2 + \sigma^2 = ECM_{TR}$$

Es decir, el cuadrado medio de los rebaños estima una componente más (la debida a los rebaños) que el cuadrado medio de la interacción, luego, si se dividen ambos cuadrados medios, se espera un valor de  $F$  significativo, esto es, estadísticamente superior a uno.

Por lo que el estadístico de prueba para el segundo factor será

$$F_o = \frac{CM_{Rebaño}}{CM_{TR}}$$

En el caso de la prueba de la tercera fuente de variación, esta es, la de la interacción, si  $\sigma_{TR}^2 = 0$

$$ECM_{TR} = n\sigma_{TR}^2 + \sigma^2 = n0^2 + \sigma^2 = ECM_{Error}$$

Es decir, el cuadrado medio de la interacción estima lo mismo que el cuadrado medio del error, luego, si se dividen ambos cuadrados medios, se espera un valor de  $F$  no significativo, esto es, estadísticamente igual a uno.

Y si  $\sigma_{TR}^2 > 0$

$$ECM_{TR} = n\sigma_{TR}^2 + \sigma^2 > \sigma^2 = ECM_{Error}$$

Es decir, el cuadrado medio de la interacción estima una componente más (la debida a la interacción) que el cuadrado medio del error, luego, si se dividen ambos cuadrados medios, se espera un valor de  $F$  significativo, esto es, estadísticamente superior a uno.

Por lo que el estadístico de prueba para la tercera fuente de variación o interacción es

$$F_o = \frac{CM_{TR}}{CM_{Error}}$$

Quedando el análisis de esta manera

**Modelo Aleatorio**

<i>FV</i>	<i>gl</i>	<i>SC</i>	<i>CM</i>	<i>F<sub>o</sub></i>
<i>Trata</i>	<i>t-1</i>	$\frac{1}{m} \sum_i X_{i..}^2 - \frac{X_{...}^2}{N}$	$\frac{SC_{Tra}}{t-1}$	$\frac{CM_{Tra}}{CM_{TR}}$
<i>Rebaño</i>	<i>r-1</i>	$\frac{1}{m} \sum_j X_{.j.}^2 - \frac{X_{...}^2}{N}$	$\frac{SC_{Reb}}{r-1}$	$\frac{CM_{Reb}}{CM_{TR}}$
<i>TxR</i>	<i>(t-1)(r-1)</i>	$\frac{1}{n} \sum_{ij} X_{ij}^2 - \frac{1}{m} \sum_i X_{i..}^2 - \frac{1}{m} \sum_j X_{.j.}^2 + \frac{X_{...}^2}{N}$	$\frac{SCTR}{(t-1)(r-1)}$	$\frac{CM_{TR}}{CM_{Err}}$
<i>Error</i>	<i>N-tr</i>	$\sum_{ijk} X_{ijk}^2 - \frac{1}{n} \sum_{ij} X_{ij.}^2$	$\frac{SC_{Err}}{N-tr}$	
<i>total</i>	<i>N-1</i>	$\sum_{ijk} X_{ijk}^2 - \frac{X_{...}^2}{N}$		

Veamos ahora el Modelo Mixto.

En el caso de la prueba del primer factor (el fijo), si  $\sigma^2_{T=0}$

$$ECM_{Trata} = m \sigma_T^2 + n \sigma_{TR}^2 + \sigma^2 = m 0^2 + n \sigma_{TR}^2 + \sigma^2 = n \sigma_{TR}^2 + \sigma^2 = ECM_{TR}$$

Es decir, el cuadrado medio del primer factor estima lo mismo que el cuadrado medio de la interacción, luego, si se dividen ambos cuadrados medios, se espera un valor de *F* no significativo, esto es, estadísticamente igual a uno.

Y si  $\sigma^2_T > 0$

$$ECM_{Trata} = m \sigma_T^2 + n \sigma_{TR}^2 + \sigma^2 > n \sigma_{TR}^2 + \sigma^2 = ECM_{TR}$$

Es decir, el cuadrado medio de los tratamientos estima una componente más (la debida a los tratamientos) que el cuadrado medio de la interacción, luego, si se dividen ambos cuadrados medios, se espera un valor de *F* significativo, esto es, estadísticamente superior a uno.

Por lo que el estadístico de prueba para el primer factor será

$$F_o = \frac{CM_{Trata}}{CM_{TR}}$$

En el caso de la prueba del segundo factor (el aleatorio), si  $\sigma_R^2 = 0$

$$E CM_{Rebaño} = tn \sigma_R^2 + \sigma^2 = tn 0^2 + \sigma^2 = \sigma^2 = E CM_{Error}$$

Es decir, el cuadrado medio de los rebaños estima lo mismo que el cuadrado medio del error, luego, si se dividen ambos cuadrados medios, se espera un valor de  $F$  no significativo, esto es, estadísticamente igual a uno.

Y si  $\sigma_R^2 > 0$

$$E CM_{Rebaño} = tn \sigma_R^2 + \sigma^2 > \sigma^2 = E CM_{Error}$$

Es decir, el cuadrado medio de los rebaños estima una componente más (la debida a los rebaños) que el cuadrado medio del error, luego, si se dividen ambos cuadrados medios, se espera un valor de  $F$  significativo, esto es, estadísticamente superior a uno.

Por lo que el estadístico de prueba para el segundo factor será

$$F_o = \frac{CM_{Rebaño}}{CM_{Error}}$$

En el caso de la prueba de la tercera fuente de variación, esta es, la de la interacción, si  $\sigma_{TR}^2 = 0$

$$E CM_{TR} = n \sigma_{TR}^2 + \sigma^2 = n 0^2 + \sigma^2 = E CM_{Error}$$

Es decir, el cuadrado medio de la interacción estima lo mismo que el cuadrado medio del error, luego, si se dividen ambos cuadrados medios, se espera un valor de  $F$  no significativo, esto es, estadísticamente igual a uno.

Y si  $\sigma_{TR}^2 > 0$

$$E CM_{TR} = n \sigma_{TR}^2 + \sigma^2 > \sigma^2 = E CM_{Error}$$

Es decir, el cuadrado medio de la interacción estima una componente más (la debida a la interacción) que el cuadrado medio del error, luego, si se dividen ambos cuadrados medios, se espera un valor de  $F$  significativo, esto es, estadísticamente superior a uno.

Por lo que el estadístico de prueba para la tercera fuente de variación o interacción es

$$F_o = \frac{CM_{TR}}{CM_{Error}}$$

Quedando el análisis de esta manera

**Modelo Mixto**

FV	gl	SC	CM	F <sub>o</sub>
Trata	t-1	$\frac{1}{m} \sum_i X_{i..}^2 - \frac{X_{...}^2}{N}$	$\frac{SC_{Tra}}{t-1}$	$\frac{CM_{Tra}}{CM_{TR}}$
Rebañ	r-1	$\frac{1}{tn} \sum_j X_{.j.}^2 - \frac{X_{...}^2}{N}$	$\frac{SC_{Reb}}{r-1}$	$\frac{CM_{Reb}}{CM_{Err}}$
TxR	(t-1) (r-1)	$\frac{1}{n} \sum_{ij} X_{ij}^2 - \frac{1}{m} \sum_i X_{i..}^2 - \frac{1}{tn} \sum_j X_{.j.}^2 + \frac{X_{...}^2}{N}$	$\frac{SC_{TR}}{(t-1)(r-1)}$	$\frac{CM_{TR}}{CM_{Err}}$
Error	N-tr	$\sum_{ijk} X_{ij}^2 - \frac{1}{n} \sum_{ij} X_{ij}^2$	$\frac{SC_{Err}}{N-tr}$	
total	N-1	$\sum_{ijk} X_{ijk}^2 - \frac{X_{...}^2}{N}$		

**Ejemplo.-**

Resolvamos, ahora, los tres enunciados planteados en el epígrafe *Factores de efectos fijos y factores de efectos aleatorios* y los tres epígrafes siguientes. Los datos para los tres enunciados eran los mismo, estos son

	T <sub>1</sub>	T <sub>2</sub>	T <sub>3</sub>	
R <sub>1</sub>	5	6	4	29
	4	7	3	
R <sub>2</sub>	1	2	1	7
	2	1	0	
	12	16	8	36

El cálculo de los resultados comunes para los tres posibles análisis son

$$\sum T^2 = \frac{1}{m} \sum_i X_{i..}^2 = \frac{1}{4} (12^2 + 16^2 + 8^2) = 116$$

$$\sum R^2 = \frac{1}{tn} \sum_j X_{.j.}^2 = \frac{1}{6} (29^2 + 7^2) = 148.3333$$

$$\sum TR^2 = \frac{1}{n} \sum_{ij} X_{ij.}^2 = \frac{1}{2} (9^2 + 13^2 + 7^2 + 3^2 + 3^2 + 1^2) = 159$$

$$TC = \frac{X_{...}^2}{N} = \frac{36^2}{3 \times 2 \times 4} = 108$$

$$\sum X^2 = \sum_{ij} X_{ij}^2 = 5^2 + 4^2 + \dots + 1^2 + 0^2 = 162$$

Enunciado 1: En este enunciado, el primer factor se trata de tres *tratamientos* que podían ser tres dietas o tres antibióticos o tres tóxicos o tres métodos de purificación, etc., es decir, un factor con tres únicos niveles del cual queremos saber si existe diferencias entre dichos niveles y en caso positivo, saber cuál o cuales es el mejor. Con respecto al segundo factor, *raza*, ocurre otro tanto, son dos razas y queremos saber si son diferentes y en caso positivo saber cual es la mejor. Por lo tanto los efectos de los tratamientos son *fijos*. No hay otros efectos pues son los únicos tratamientos que interesan. Lo mismo para las dos razas. El análisis es

#### Modelo Fijo

FV	gl	SC	CM	ECM	F <sub>o</sub>
Trata	3-1	116-108	$\frac{8}{2}=4$	$\sigma^2 + 4\sigma_T^2$	$\frac{4}{0.50}=8.00^*$
Raza	2-1	148.33-108	$\frac{40.33}{1}=40.33$	$\sigma^2 + 6\sigma_R^2$	$\frac{40.33}{0.50}=80.66^{***}$
TxR	(3-1) (2-1)	159-116- 148.33+108	$\frac{2.66}{2}=1.33$	$\sigma^2 + 2\sigma_{TR}^2$	$\frac{1.33}{0.50}=2.66ns$
Error	12-6	162-159	$\frac{3}{6}=0.50$	$\sigma^2$	
total	12-1	162-108			

Las F de contraste son, para tratamiento  $F_{(2,6; 0.05)} = 5.14$ , para raza  $F_{(1,6; 0.05)} = 5.9$  para la interacción  $F_{(2,6; 0.05)} = 5.14$

Como vemos, los dos factores principales (tratamiento y raza) son significativos, mientras que la interacción es no significativa. Por lo que se concluye que hay diferencia entre los efectos de los tratamientos; al menos hay dos de ellos diferentes, es decir, hay una o unas dietas mejor que otra u otras. También hay diferencia entre los efectos de las dos razas; una de ellas es superior a la otra. Sin embargo no existe

interacción, esto es, el orden de magnitud de los efectos de un factor se mantiene dentro de los diferentes niveles del otro factor.

Enunciado 2: Supóngase se nos hubiera informado que la experiencia se realizó con animales de dos rebaños diferentes, de manera que el primer factor se trata de tres *tratamientos* (como en el enunciado anterior) del cual queremos saber si existe diferencias entre dichos tratamientos y en caso positivo, saber cuál o cuales es el mejor. Con respecto al segundo factor, *rebaño*, no ocurre lo mismo sino que este segundo factor lo ha introducido el experimentador debido a criterios ambientales o de clasificación. El experimentador no tiene interés en saber si existen diferencias entre los niveles de este segundo factor, de hecho se introduce esta segunda fuente de variación porque se sabe, con cierta certeza, que influye en la variabilidad de la variable que se está midiendo, es decir, se sabe que unos rebaños son diferentes a otros. El que un animal pertenezca a un rebaño o a otro no es aleatorio mientras que los tratamientos han sido aplicados aleatoriamente a cada animal dentro de cada rebaño. Lo mismo que se ha dicho de los rebaños se puede decir de las plantas de los hospitales o de los hospitales o de los barrios o de las ciudades o de las altitudes, etc.

Los rebaños son de efectos aleatorios porque al ser una muestra de los muchos otros rebaños que se podrían haber muestreado, estos efectos hubieran sido diferentes si se hubieran muestreado otros rebaños.

Como en el modelo que explica los datos de este ejemplo, uno de los factores es de efectos fijos y el otro de efectos aleatorios, se tiene un *Modelo mixto*.

Modelo Mixto

<i>FV</i>	<i>gl</i>	<i>SC</i>	<i>CM</i>	<i>ECM</i>	<i>F<sub>o</sub></i>
<i>Trata</i>	3-1	116-108	$\frac{8}{2}=4$	$\sigma^2 + 2\sigma_{TR}^2 + 4\sigma_T^2$	$\frac{4}{1.33}=3.01ns$
<i>Rebaño</i>	2-1	148.33-108	$\frac{40.33}{1}=40.33$	$\sigma^2 + 6\sigma_R^2$	$\frac{40.33}{0.50}=80.66***$
<i>TxR</i>	(3-1) (2-1)	159-116- 148.33+108	$\frac{2.66}{2}=1.33$	$\sigma^2 + 2\sigma_{TR}^2$	$\frac{1.33}{0.50}=2.66ns$
<i>Error</i>	12-6	162-159	$\frac{3}{6}=0.50$	$\sigma^2$	
<i>total</i>	12-1	162-108			

Las *F* de contraste son, para tratamiento  $F_{(2,2; 0.05)} = 19.0$ , para raza  $F_{(1,6; 0.05)} = 5.9$  para la interacción  $F_{(2,6; 0.05)} = 5.14$

Como se ve, en este caso es significativo solo el factor rebaño, siendo no significativo los tratamientos y la interacción. Por lo que se concluye que las tres dietas



son iguales y no interaccionan con los rebaños, es decir, que son iguales en cualquier rebaño. Sin embargo, lo que es de esperar, el efecto rebaño es significativo.

Enunciado 3: En los enunciados 1 y 2 los tres *Tratamientos* eran fijos, por ejemplo, tres dietas o tres antibióticos o tres tóxicos, etc., es decir, un factor con tres únicos niveles del cual queremos saber si existe diferencias entre ellos y cuál es el mejor. El segundo factor del primer enunciado también era fijo, eran dos razas y se quería saber cuál era la mejor. Pero el segundo factor del segundo enunciado era aleatorio, los dos niveles muestreados es una muestra aleatoria de los muchos niveles (rebaños) existentes de dicho factor. Ahora se va a suponer que los tres niveles del primer factor también van ha ser aleatorios, es decir, una muestra aleatoria de los muchos niveles existentes. Por ejemplo, supóngase que se esta estudiando la resistencia o capacidad de asimilación del ganado vacuno lechero, de cierto tóxico natural que se produce espontáneamente en el almacenamiento de cierto pienso, para ello se toman cuatro *medioshermanos* hijos de cada uno de tres *machos* repartidos dos a dos en dos ganaderías y se le analiza la concentración de dicho tóxico en el hígado. Por tanto, en este caso los tres tratamientos o tres niveles del primer factor son las tres familias, que es una muestra aleatoria de las muchas familias que se podrían haber utilizado. Siendo los rebaños, como en el segundo enunciado, una muestra aleatoria de los muchos rebaños que se podrían haber utilizado.

Como en el modelo que explica los datos de este ejemplo, los dos factores son de efectos aleatorios, se dice que es un *Modelo aleatorio* o *Modelo II*.

Modelo Aleatorio

<i>FV</i>	<i>gl</i>	<i>SC</i>	<i>CM</i>	<i>ECM</i>	<i>F<sub>o</sub></i>
<i>Fami</i>	3-1	116-108	$\frac{8}{2}=4$	$\sigma^2 + 2\sigma_{TR}^2 + 4\sigma_T^2$	$\frac{4}{1.33}=3.01ns$
<i>Rebaño</i>	2-1	148.33-108	$\frac{40.33}{1}=40.33$	$\sigma^2 + 2\sigma_{TR}^2 + 6\sigma_R^2$	$\frac{40.33}{1.33}=30.25 *$
<i>TxR</i>	(3-1) (2-1)	159-116- 148.33+108	$\frac{2.66}{2}=1.33$	$\sigma^2 + 2\sigma_{TR}^2$	$\frac{1.33}{0.50}=2.66ns$
<i>Error</i>	12-6	162-159	$\frac{3}{6}=0.50$	$\sigma^2$	
<i>total</i>	12-1	162-108			

Las *F* de contraste son, para tratamiento  $F_{(2,2; 0.05)} = 19.00$ , para raza  $F_{(1,2; 0.05)} = 18.5$  para la interacción  $F_{(2,6; 0.05)} = 5.14$

Como se ve es significativo solo el factor rebaño, no lo es ni los tratamientos (familias) ni existe interacción. Por lo que se concluye que la capacidad de asimilación

de dicho tóxico viene influida por el factor rebaño, no influye el factor familia ni existe interacción familia/rebaño.

Veamos como se puede resolver estos tres posibles enunciados con el SAS

**Archivo del programa SAS (C7-1.SAS).-**

```

title 'Análisis de la varianza, Factorial con repetición';
option ps=65 ls=75;
data facta;
infile 'C7-1.dat';
input trata $ raza $ prod @@;
title 'Modelo Fijo';
proc ANOVA;
  class trata raza;
  model prod=trata raza trata*raza;
run;
title 'Modelo Mixto (trata fijo)';
  test H=trata E=trata*raza;
run;
title 'Modelo Aleatorio';
proc glm;
  class trata raza;
  model prod = trata raza trata*raza;
  random trata raza trata*raza / test ;
run;

```

**Archivo de datos (C7-1.DAT).-**

T1 R1 5	T2 R1 6	T3 R1 3
T1 R1 4	T2 R1 7	T3 R1 4
T1 R2 1	T2 R2 2	T3 R2 1
T1 R2 2	T2 R2 1	T3 R2 0

**Archivo de resultados (C7-1.LST).-**

Modelo Fijo					
Dependent Variable: PROD					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	51.0000000	10.2000000	20.40	0.0011
Error	6	3.0000000	0.5000000		
Corrected Total	11	54.0000000			
	R-Square	C.V.	Root MSE	PROD Mean	
	0.944444	23.57023	0.70711	3.00000	
Source	DF	Anova SS	Mean Square	F Value	Pr > F
TRATA	2	8.0000000	4.0000000	8.00	0.0203
RAZA	1	40.3333333	40.3333333	80.67	0.0001
TRATA*RAZA	2	2.6666667	1.3333333	2.67	0.1484
Modelo Mixto (trata fijo)					
Tests of Hypotheses using the Anova MS for TRATA*RAZA as an error term					
Source	DF	Anova SS	Mean Square	F Value	Pr > F
TRATA	2	8.0000000	4.0000000	3.00	0.2500

Modelo Aleatorio

Source            Type III Expected Mean Square  
 TRATA            Var(Error) + 2 Var(TRATA\*RAZA) + 4 Var(TRATA)  
 RAZA             Var(Error) + 2 Var(TRATA\*RAZA) + 6 Var(RAZA)  
 TRATA\*RAZA     Var(Error) + 2 Var(TRATA\*RAZA)

Tests of Hypotheses for Random Model Analysis of Variance

Dependent Variable: PROD  
 Source: TRATA  
 Error: MS(TRATA\*RAZA)

DF	Type III MS	Denominator DF	Denominator MS	F Value	Pr > F
2	4	2	1.3333333333	3.0000	0.2500

Source: RAZA  
 Error: MS(TRATA\*RAZA)

DF	Type III MS	Denominator DF	Denominator MS	F Value	Pr > F
1	40.333333333	2	1.3333333333	30.2500	0.0315

Source: TRATA\*RAZA  
 Error: MS(Error)

DF	Type III MS	Denominator DF	Denominator MS	F Value	Pr > F
2	1.3333333333	6	0.5	2.6667	0.1484

**Experimentos factoriales.-**

En todos los modelos vistos hasta ahora con dos factores, el segundo factor venia impuesto por motivos ambientales o de clasificación. El interés de los problemas eran los tres tratamientos, que podían ser tres dietas o tres antibióticos o tres procesos de laboratorio o industriales, y se introducía un segundo factor, impuesto por las condiciones experimentales, con objeto de disminuir la varianza del error y, de esta manera, aumentar la precisión del análisis .

Pero podría ocurrir que hubiera un segundo factor porque el experimentador este interesado en los diferentes niveles de este factor como otros tratamientos; por ejemplo, se podría estar interesado en estudiar tres concentraciones de vitaminas y dos racionamientos, o tres antibióticos y dos vacunas, o tres procesos de elaboración de cierto alimento precocinado y dos aditivos, etc.

En casos como estos puede haber dos estrategias de estudio: la de realizar un experimento para cada factor o la de estudiar los dos factores en un mismo experimento. Esta segunda estrategia tiene ventajas económicas y científicas con respecto a la primera.

**Ejemplo.-**

Se quiere probar tres tratamientos, por ejemplo tres concentraciones de vitaminas, y además se quiere probar dos racionamientos, por lo que en total se tiene seis tratamientos: los tres para las tres concentraciones de vitaminas con el primer racionamiento y los tres de las tres concentraciones de vitaminas con el segundo racionamiento. O lo que es lo mismo, los tratamientos son la primera concentración de vitamina con los dos racionamientos, la segunda concentración de vitamina con los dos racionamientos y la tercera concentración de vitamina con los dos racionamientos. Si expresamos estos tratamientos en una tabla de datos se tendría

$R_1 V_1$	$R_1 V_2$	$R_1 V_3$	$R_2 V_1$	$R_2 V_2$	$R_2 V_3$
5	6	4	1	2	1
4	7	3	2	1	0

Esta tabla se podría organizar de esta otra manera

	$V_1$	$V_2$	$V_3$	
$R_1$	5 4	6 7	4 3	29
$R_2$	1 2	2 1	1 0	7
	12	16	8	36

Como se ve, este ejemplo es, formalmente, igual que el ejemplo utilizado anteriormente para los modelos factoriales con repetición, aunque conceptualmente es diferente. Es decir, en el ejemplo anterior, el que una unidad experimental pertenezca a una  $R$  o a otra no era aleatorio, sino que era el experimentador el que lo determinaba con arreglo a criterios ajenos al problema, normalmente criterios de clasificación. Aquí, la adjudicación tanto de las concentraciones de vitamina como de los racionamientos, son aleatorias. En este ejemplo se ha tenido que tomar doce unidades experimentales, numerarlas del uno al doce, elegir dos al azar y adjudicarle el primer tratamiento, elegir otras dos y adjudicarle el segundo tratamiento, y así hasta que queden dos unidades experimentales que serán a las que se le aplique el sexto tratamiento.

Aunque conceptualmente es diferente a los anteriores ejemplos factoriales, formalmente es lo mismo, por lo que también es lo mismo lo dicho con respecto al tipo de factor (fijo o aleatorio) y con respecto a la interacción. Por tanto el modelo lineal que explica esta experiencia es el mismo que el anterior

$$X_{ijk} = \mu + V_i + R_j + VR_{ij} + e_{ijk}$$

La resolución de este ejemplo puede ser la misma que la utilizada con los diseños factoriales con repetición o puede hacerse como si fuera un solo factor con seis tratamientos.

Más adelante se amplía el estudio de los experimentos factoriales y la derivaciones más importantes.

**Generalización del método para el análisis de modelos factoriales con repetición con más de dos factores.-**

Es importante conocer la mecánica del cálculo y análisis que debe seguirse para diversos tipos de modelos, pues aunque hasta ahora se han mostrado sus expresiones para el casos de factoriales con dos factores, interesa saber la generalización a cualquier número de factores.

En el caso de que se tenga un modelo factorial con  $f$  factores, el modelo sería

$$\begin{aligned}
 X_{ijk\dots f} = & \mu + A_i + B_j + C_k + \dots + AB_{ij} + AC_{ik} + AD_{il} + \dots + \\
 & + ABC_{ijk} + ABD_{ijl} + \dots + ABCD_{ijkl} + ABCE_{ijkm} + \dots + \\
 & + \dots\dots\dots + \\
 & + e_{ijk\dots f}
 \end{aligned}$$

Los datos habría que ordenarlos en tablas de doble entrada de manera que, por ejemplo, para un factorial triple con dos niveles en  $A$  y en  $C$  y tres en  $B$  se tendría dos tablas de datos de la siguiente manera

A <sub>1</sub>				A <sub>2</sub>				
				B <sub>1</sub>	B <sub>2</sub>	B <sub>3</sub>		
C <sub>1</sub>	X <sub>1111</sub>	X <sub>1211</sub>	X <sub>1311</sub>	X <sub>2111</sub>	X <sub>2211</sub>	X <sub>2311</sub>	X <sub>..1</sub>	
	X <sub>1112</sub>	X <sub>1212</sub>	X <sub>1312</sub>	X <sub>2112</sub>	X <sub>2212</sub>	X <sub>2312</sub>		
C <sub>2</sub>	X <sub>1121</sub>	X <sub>1221</sub>	X <sub>1321</sub>	X <sub>2121</sub>	X <sub>2221</sub>	X <sub>2321</sub>	X <sub>..2</sub>	
	X <sub>1122</sub>	X <sub>1222</sub>	X <sub>1322</sub>	X <sub>2122</sub>	X <sub>2222</sub>	X <sub>2322</sub>		
X <sub>1...</sub>				X <sub>1..</sub>	X <sub>2..</sub>	X <sub>3..</sub>	X <sub>2...</sub>	

O bien los podríamos ordenar en una lista de  $A \times B \times C$  filas y  $f+n$  columnas, de manera que los datos de las  $f$  primeras columnas hicieran referencia el nivel de cada uno de los  $f$  factores y los  $X$  últimos datos se refiera al valor de la variable en cada casilla. Para el mismo ejemplo de las tablas anteriores tendríamos algo parecido a esto

A <sub>1</sub>	B <sub>1</sub>	C <sub>1</sub>	X <sub>1111</sub>	X <sub>1112</sub>	A <sub>2</sub>	B <sub>1</sub>	C <sub>1</sub>	X <sub>2111</sub>	X <sub>2112</sub>
A <sub>1</sub>	B <sub>1</sub>	C <sub>2</sub>	X <sub>1121</sub>	X <sub>1122</sub>	A <sub>2</sub>	B <sub>1</sub>	C <sub>2</sub>	X <sub>2121</sub>	X <sub>2122</sub>
A <sub>1</sub>	B <sub>1</sub>	C <sub>3</sub>	X <sub>1131</sub>	X <sub>1132</sub>	A <sub>2</sub>	B <sub>1</sub>	C <sub>3</sub>	X <sub>2131</sub>	X <sub>2132</sub>
A <sub>1</sub>	B <sub>2</sub>	C <sub>1</sub>	X <sub>1211</sub>	X <sub>1212</sub>	A <sub>2</sub>	B <sub>2</sub>	C <sub>1</sub>	X <sub>2211</sub>	X <sub>2212</sub>
A <sub>1</sub>	B <sub>2</sub>	C <sub>2</sub>	X <sub>1221</sub>	X <sub>1222</sub>	A <sub>2</sub>	B <sub>2</sub>	C <sub>2</sub>	X <sub>2221</sub>	X <sub>2222</sub>
A <sub>1</sub>	B <sub>2</sub>	C <sub>3</sub>	X <sub>1231</sub>	X <sub>1232</sub>	A <sub>2</sub>	B <sub>2</sub>	C <sub>3</sub>	X <sub>2231</sub>	X <sub>2232</sub>

Como vemos,  $B_1$  en  $A_1$  es el mismo que en  $A_2$ , y es el mismo en los tres niveles de  $C$ , etc.

El modelo lineal que explicaría esta experiencia sería

$$X_{ijkl} = \mu + A_i + B_j + C_k + AB_{ij} + AC_{ik} + BC_{jk} + ABC_{ijk} + e_{ijkl}$$

El sistema para calcular los  $gl$  y  $SC$  tiene una lógica común; tan sólo hay que tener en cuenta que cuando se quiera hallar los grados de libertad los números serán los niveles de los factores y para las sumas de cuadrados serán a los sumatorios de los niveles correspondientes. Con el fin de hacer la exposición más fácil vamos a llamar a los sumatorios de los factores  $A, B, C$ , etc. como  $\Sigma A^2, \Sigma B^2, \Sigma C^2$ , etc., siendo su expresión matemática

$$\Sigma A^2 = \Sigma_i \frac{X_{i..}^2}{bcn}; \quad \Sigma B^2 = \Sigma_j \frac{X_{.j.}^2}{acn}; \quad \Sigma C^2 = \Sigma_k \frac{X_{..k}^2}{abn}$$

Los grados de libertad de los factores, al igual que en el modelo de una vía, se obtienen restandole la unidad al número de niveles del factor considerado:  $a-1, b-1, c-1, \dots, f-1$ .

Para la suma de cuadrados se resta del sumatorio del factor el término de corrección:  $\Sigma A^2 - TC, \Sigma B^2 - TC, \Sigma C^2 - TC$ .

Los valores para las interacciones se pueden obtener de dos formas

a) Para los grados de libertad, multiplicando los  $gl$  de los factores integrantes  $(a-1)(b-1)$ .

En el caso de las  $SC$ , por una multiplicación simbólica  $SC(AB) = (\Sigma A^2 - 1)(\Sigma B^2 - 1) = \Sigma AB^2 - \Sigma A^2 - \Sigma B^2 + TC$ , donde hemos sustituido el uno por el  $TC$ .

b) Para los  $gl$ , restando del total de niveles de la interacción  $(axb)$  los de cada uno de los factores y añadiendo luego la unidad:  $ab - a - b + 1$ .

Paralelamente, la  $SC$  sería el sumatorio de la interacción a estudiar, menos los sumatorios de los factores simples integrantes, más el  $TC$ :  $SC(AB) = \Sigma AB^2 - \Sigma A^2 - \Sigma B^2 + TC$ . Es decir, a la interacción  $AB$  hay que restarle lo debido a  $A$  y  $B$  pero como ambos llevan implícito lo debido a la corrección de la media general,  $TC$ , habrá que sumarle una vez este término ya que se ha restado dos veces.

En general, en los modelos factoriales puros, los valores para las interacciones múltiples se obtienen, tanto para la  $SC$  como para los  $gl$ , restando de la interacción problema todas las del orden inmediato inferior (con un factor menos), sumando luego todas las del orden siguiente (otro factor menos), restando las del orden siguiente, y así sucesivamente, alternando los signos, hasta llegar al  $TC$ , en la  $SC$ , o al uno en los  $gl$ .

Resumiendo lo dicho hasta ahora en un cuadro, se tiene

FV	gl	SC
A	a-1	$\sum_i \frac{X_{i...}^2}{bcn} - \frac{X^2}{N}$
B	b-1	$\sum_j \frac{X_{.j.}^2}{acn} - \frac{X^2}{N}$
C	c-1	$\sum_k \frac{X_{..k.}^2}{abn} - \frac{X^2}{N}$
AB	(a-1)(b-1)= ab-a-b+1	$\sum_{ij} \frac{X_{ij.}^2}{cn} - \sum_i \frac{X_{i...}^2}{bcn} - \sum_j \frac{X_{.j.}^2}{acn} + \frac{X^2}{N}$
AC	ac-a-c+1	$\sum_{ik} \frac{X_{i.k.}^2}{bn} - \sum_i \frac{X_{i...}^2}{bcn} - \sum_k \frac{X_{..k.}^2}{abn} + \frac{X^2}{N}$
BC	bc-b-c+1	$\sum_{jk} \frac{X_{.jk.}^2}{an} - \sum_j \frac{X_{.j.}^2}{acn} - \sum_k \frac{X_{..k.}^2}{abn} + \frac{X^2}{N}$
ABC	(a-1)(b-1)(c-1)= =abc-ab-ac- bc+a+b+c-1	$\sum_{ijk} \frac{X_{ijk.}^2}{n} - \sum_{ij} \frac{X_{ij.}^2}{cn} - \sum_{ik} \frac{X_{i.k.}^2}{bn} - \sum_{jk} \frac{X_{.jk.}^2}{an} +$ $\sum_i \frac{X_{i...}^2}{bcn} + \sum_j \frac{X_{.j.}^2}{acn} + \sum_k \frac{X_{..k.}^2}{abn} - \frac{X^2}{N}$
Error	N-abc	$\sum_{ijkl} X_{ijkl}^2 - \sum_{ijk} \frac{X_{ijk.}^2}{n}$
total	abc-1=N-1	$\sum_{ijkl} X_{ijkl}^2 - \frac{X^2}{N}$

Una vez hallados los grados de libertad y las sumas de cuadrados, los cuadrados medios son las divisiones respectivas de las sumas de cuadrados por los grados de libertad, por lo que se esta en condiciones de probar las diferentes fuentes de variación. Para ello se tiene que saber cuales son los parámetros que estiman los cuadrados medios y, como se sabe, estos cambian con arreglo al tipo de factores que se tengan (fijos o aleatorios), por lo que se necesita un método que nos provea de estas estimas para saber, entre otras finalidades, cuales son las pruebas de significación para las diferentes fuentes de variación. Este método se verá más adelante, en el epígrafe *Cálculo de las estimas de los CM por el método rápido*.

Solamente si todos los factores son de efectos fijos se sabe que cada cuadrado medio estima, además de la varianza de su fuente de variación, la varianza del error, por lo que todas las pruebas de hipótesis usarían como termino de error el cuadrado medio del *Error*.

**Ejemplo.-**

Se quiere saber las incidencias que tienen dos dietas en la producción de leche en cabras. Como se sabe que, en dicha producción, también influye la edad de la cabra o número de lactaciones que haya tenido y la estación del año en que la cabra comienza la lactación, se hace la experiencia con cabras de primera, segunda y tercera lactación y en las dos estaciones del año habituales (otoño, invierno). Siendo los datos

$T_1$				$T_2$					
	$L_1$	$L_2$	$L_3$		$L_1$	$L_2$	$L_3$		
O	325	374	364	O	375	401	392		
	345	395	359		384	395	385		
I	304	352	336	I	352	395	362		
	310	348	328		349	360	365		
4140				2744			3020	2891	4515

**Archivo del programa SAS (C7-2.SAS) .-**

```
Title 'Análisis de la varianza, Factorial con tres factores';
option ps=65 ls=75;
data fact3;
infile 'C7-2.dat';
input trata $ nlac $ epoca $ prod @@;
proc ANOVA;
class trata nlac epoca;
model prod = trata nlac epoca trata*nlac trata*epoca
nlac*epoca trata*nlac*epoca;
run;
```

**Archivo de datos (C7-2.DAT) .-**

```
T1 L1 O 325 T1 L2 O 374 T1 L3 O 364
T1 L1 O 345 T1 L2 O 395 T1 L3 O 359
T1 L1 I 304 T1 L2 I 352 T1 L3 I 336
T1 L1 I 310 T1 L2 I 348 T1 L3 I 328
T2 L1 O 375 T2 L2 O 401 T2 L3 O 392
T2 L1 O 384 T2 L2 O 395 T2 L3 O 385
T2 L1 I 352 T2 L2 I 395 T2 L3 I 362
T2 L1 I 349 T2 L2 I 360 T2 L3 I 365
```

**Archivo de resultados (C7-2.LST) .-**

Dependent Variable: PROD						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	11	15922.1250	1447.4659	14.53	0.0001	
Error	12	1195.5000	99.6250			
Corrected Total	23	17117.6250				
	R-Square	C.V.	Root MSE	PROD Mean		
	0.930160	2.767759	9.98123	360.625		



Source	DF	Anova SS	Mean Square	F Value	Pr > F
TRATA	1	5859.37500	5859.37500	58.81	0.0001
NLAC	2	4767.75000	2383.87500	23.93	0.0001
EPOCA	1	4620.37500	4620.37500	46.38	0.0001
TRATA*NLAC	2	564.25000	282.12500	2.83	0.0983
TRATA*EPOCA	1	51.04167	51.04167	0.51	0.4878
NLAC*EPOCA	2	1.75000	0.87500	0.01	0.9913
TRATA*NLAC*EPOCA	2	57.58333	28.79167	0.29	0.7541

Por lo que se concluye que los efectos de los tres factores principales son significativos. Ninguna interacción influye en la variabilidad de los datos.

Recuérdese que este es un caso particular en el que todos los factores son de efectos fijos, por lo que el término de error, en todas las pruebas, es el cuadrado medio del *Error*, esto no será así si algún o algunos factores hubieran sido de efectos aleatorios, en cuyo caso habría que haber hallado las estimas de los cuadrados medios para saber cuales son los términos de error de las diferentes pruebas. Cómo se hallan estas estimas se verá en el epígrafe *Cálculo de las estimas de los CM por el método rápido*, y en el epígrafe *Prueba F aproximada* se verá como se resuelve este mismo ejemplo para el caso en el que los tratamientos sean fijos y los otros dos factores sean de efectos aleatorios.

### Un factorial con repetición en un factorial sin repetición.-

Un tipo de modelo con tres factores que se puede presentar con cierta frecuencia es el consistente en un experimento factorial con repetición de dos factores, pero en el que las repeticiones se han realizado de manera tal que puede considerarse tengan un efecto ambiental, por lo que hay que introducirlas como un factor más, de manera que, como las repeticiones se introducen como un factor, se tiene, con arreglo a este factor, un factorial sin repetición. Este sería un experimento factorial en un diseño de bloques aleatorios, un ejemplo sería el siguiente.

### Ejemplo.-

Se quiere saber cuál es el incremento de peso, en cierta especie cárnica, en función de dos raciones ( $a_1$  y  $a_2$ ) y los dos sexos ( $M$  y  $F$ ). Los tratamientos serán, por tanto,  $a_1F$ ,  $a_1M$ ,  $a_2F$  y  $a_2M$ . Se hicieron cuatro repeticiones en sendas *jaulas*, que constituían, por tanto cuatro bloques. Los datos son

			Ración		
			$a_1$	$a_2$	
Jaula	1	sexo	M	9.52	8.51
			F	9.94	10.00
	2	sexo	M	8.21	9.95
			F	9.48	9.24
	3	sexo	M	9.32	8.43
			F	9.32	9.34
	4	sexo	M	10.56	8.86
			F	10.90	9.68

Esta claro que el modelo lineal que explica esta experiencia es

$$X_{ijk} = \mu + J_i + S_j + R_k + SR_{jk} + e_{ijk}$$

El factor jaula es de efectos aleatorios y los demás son de efectos fijos, por lo que todos los contrastes se harán con el cuadrado medio del error.

El análisis es el siguiente.

$$\sum J^2 = \frac{1}{sr} \sum_i X_{i..}^2 = \frac{1}{2 \times 2} (37.97^2 + 36.88^2 + 36.41^2 + 40^2) = 1431.8858$$

$$\sum S^2 = \frac{1}{jr} \sum_j X_{.j.}^2 = \frac{1}{4 \times 2} (73.36^2 + 77.9^2) = 1431.2625$$

$$\sum R^2 = \frac{1}{js} \sum_k X_{...k}^2 = \frac{1}{4 \times 2} (77.25^2 + 74.01^2) = 1430.6303$$

$$\sum SR^2 = \frac{1}{j} \sum_{jk} X_{.jk}^2 = \frac{1}{4} (37.61^2 + 35.75^2 + 39.64^2 + 38.26^2) = 1431.9329$$

$$TC = \frac{X_{...}^2}{jsr} = \frac{151.26^2}{4 \times 2 \times 2} = 1429.9742$$

$$\sum X^2 = \sum_{ijk} X_{ijk}^2 = 9.52^2 + 8.51^2 + \dots + 10.90^2 + 9.68^2 = 1438.0596$$

FV	gl	SC	CM	F
<i>Jaula</i>	3	1431.8858-1429.9742	$\frac{1.9116}{3} = 0.6372$	$\frac{0.6372}{0.4683} = 1.36ns$
<i>Sexo</i>	1	1431.2624-1429.9742	$\frac{1.2883}{1} = 1.2883$	$\frac{1.2883}{0.4683} = 2.75ns$
<i>Raci</i>	1	1430.6303-1429.9742	$\frac{0.6561}{1} = 0.6561$	$\frac{0.6561}{0.4683} = 1.40ns$
<i>SR</i>	1	1431.9329-1431.2624- 1430.6303+1429.9742	$\frac{0.0144}{1} = 0.0144$	$\frac{0.01435}{0.4683} = 0.03ns$
<i>Error</i>	9	1438.0596-1431.9329- 1431.8858+1429.9742	$\frac{4.2151}{9} = 0.4683$	
<i>total</i>	15	1438.0596-1429.9742		

Las *F* de contraste son, para *Jaula*,  $F_{(3,9; 0.05)} = 3.86$ , y para las demás fuentes de variación,  $F_{(1,9; 0.05)} = 5.12$ .

Por lo que se concluye que no son significativos los efectos de ninguno de los factores, así como no es significativa la interacción. Esto es, no influye, en el aumento de peso, ni la jaula ni el sexo ni la ración. Esto no es sorprendente si no se ha tenido en cuenta los pesos iniciales, pues la variabilidad de los pesos iniciales puede haber ocultado la variabilidad debida a alguno o algunos de los factores considerados. Un análisis más correcto para este experimento hubiera sido un *análisis de covarianza* (ver Capítulo 20) de la variable aumento de peso con la covariable peso inicial. O este análisis hubiera sido más útil si la experiencia se hubiera realizado, si ello es posible, con individuos que tengan el mismo peso al comenzar la experiencia

### Archivo del programa SAS (C7-3.SAS).-

```

title 'Factorial con repetición en un factorial sin repetición';
options ls=75 ps=60;
data fac_ba;
infile 'c7-3.dat';
input jaula racion $ sexo $ aupeso @@;
proc anova;
class jaula racion sexo;
model aupeso = jaula racion sexo racion*sexo;
run;

```

### Archivo de datos (C7-3.DAT) .-

```

1 a1 M 9.52    1 a2 M 8.51
1 a1 F 9.94    1 a2 F 10.00
2 a1 M 8.21    2 a2 M 9.95
2 a1 F 9.48    2 a2 F 9.24

```

3 a1 M	9.32	3 a2 M	8.43
3 a1 F	9.32	3 a2 F	9.34
4 a1 M	10.56	4 a2 M	8.86
4 a1 F	10.90	4 a2 F	9.68

### Archivo de resultados (C7-3.LST).-

Dependent Variable: AUPESO						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	6	3.87035000	0.64505833	1.38	0.3194	
Error	9	4.21502500	0.46833611			
Corrected Total	15	8.08537500				
	R-Square	C.V.	Root MSE	AUPESO Mean		
	0.478685	7.238936	0.68435	9.45375		
Source	DF	Anova SS	Mean Square	F Value	Pr > F	
JAULA	3	1.91162500	0.63720833	1.36	0.3157	
RACION	1	0.65610000	0.65610000	1.40	0.2669	
SEXO	1	1.28822500	1.28822500	2.75	0.1316	
RACION*SEXO	1	0.01440000	0.01440000	0.03	0.8647	

### Bibliografía:

- Cochran, WG y Cox, GM.* 1965. DISEÑOS EXPERIMENTALES. Ed. TRILLAS. México.
- Li, Ching Chun.* 1969. INTRODUCCIÓN A LA ESTADÍSTICA EXPERIMENTAL. Ed. Omega. Barcelona.
- Lite, TM, y Jackson Hills, F.* 1987. MÉTODOS ESTADÍSTICOS PARA LA INVESTIGACIÓN EN LA AGRICULTURA. Ed. TRILLAS. México.
- Littel, R.C., Freund, R.J. and Spector, P.C.* 1991. SAS® FOR LINEAR MODELS. SAS Institute Inc., Cary, NC, USA.
- Ostle, B.* 1965. ESTADÍSTICA APLICADA. Ed. Limusa-Wiley. México.
- Ruiz-Maya, L.* 1983. MÉTODOS ESTADÍSTICOS DE INVESTIGACIÓN. Ed. INE. Madrid.
- Sokal, R.R. y Rohlf, F.J.* 1994. BIOMETRY. Ed. W.H.FREEMAN. San Francisco.
- Snedecor, G.W. y Cochran, W.G.* 1971. MÉTODOS ESTADÍSTICOS. Ed. C.E.C.S.A. México.
- Steel, R.* 1996. PRINCIPLES AND PROCEDURES OF STATISTICS. Ed. McGRAW-HILL Education. New York .
- SAS Institute Inc.* 1990. SAS/STAT USER'S GUIDE. Volume 1 and 2. Cary, NC, USA.



## **CAPÍTULO 8**

# **Modelos Jerárquicos**



## Modelos Jerárquicos

### Relación entre los factores.-

Cuando se tenga un modelo lineal con dos fuentes de variación no siempre se van a relacionar estas dos fuentes de variación de la manera vista en los ejemplos anteriores, es decir, *cruzándose* los niveles de los dos factores, de manera que los tres mismos tratamientos se encontraban en los dos rebaños, y en los dos rebaños se encontraban cada uno de los tres tratamientos. A los modelos cuyos factores se cruzan de la manera vista en los ejemplos anteriores se les denomina genéricamente como *Factoriales*. Otro tipo de relación que se puede presentar entre dos factores es que los niveles de un factor estén *jerarquizados* (*subordinados, encajados, anidados*) a los niveles del factor principal, de manera que dentro de un nivel del factor principal, por ejemplo el primer tratamiento, se encuentran los dos niveles del factor jerarquizado, como en el caso anterior, pero dentro de otro nivel del factor principal, por ejemplo, segundo tratamiento, se encuentran otros dos niveles del factor jerarquizado semejantes a los del primer tratamiento pero no los mismos.

### Ejemplo.-

Para continuar la explicación volvamos al primer ejemplo. Supóngase que en este caso cada uno de los tres tratamientos se hubieran aplicado en dos rebaños diferentes, de manera que se tendría los datos de la siguiente manera



$T_1$		$T_2$		$T_3$	
$R_1$	$R_2$	$R_1$	$R_2$	$R_1$	$R_2$
$X_{111} = 5$	$X_{121} = 4$	$X_{211} = 7$	$X_{221} = 6$	$X_{311} = 4$	$X_{321} = 3$
$X_{112} = 2$	$X_{122} = 1$	$X_{212} = 1$	$X_{222} = 2$	$X_{312} = 0$	$X_{322} = 1$
$X_{11.} = 7$	$X_{12.} = 5$	$X_{21.} = 8$	$X_{22.} = 8$	$X_{31.} = 4$	$X_{32.} = 4$
$\bar{X}_{11.} = 3.5$	$\bar{X}_{12.} = 2.5$	$\bar{X}_{21.} = 4$	$\bar{X}_{22.} = 4$	$\bar{X}_{31.} = 2$	$\bar{X}_{32.} = 2$
$\rho = 0.5$	$\rho = -0.5$	$\rho = 0$	$\rho = 0$	$\rho = 0$	$\rho = 0$
$X_{1..} = 12$		$X_{2..} = 16$		$X_{3..} = 8$	
$\bar{X}_{1..} = 3$		$\bar{X}_{2..} = 4$		$\bar{X}_{3..} = 2$	
$\tau_1 = 0$		$\tau_2 = 1$		$\tau_3 = -1$	

Esto indica que el  $R_1$  y el  $R_2$  del primer tratamiento son diferentes que los del segundo tratamiento y estos diferentes a los del tercer tratamiento, aunque son semejantes. Esto es, son seis rebaños repartidos dos a dos para cada tratamiento pero son rebaños que no se diferencian en ningún otro factor, es decir, son semejantes.

Si se hubiese querido poner esta tabla en forma de factores cruzados o factorial, hubiera sido de esta manera

	$T_1$	$T_2$	$T_3$
$R_1$	5 2		
$R_2$	4 1		
$R_3$		7 1	
$R_4$		6 2	
$R_5$			4 0
$R_6$			3 1

Es decir, habría casillas vacías correspondiente a los niveles del factor jerarquizado donde no se hubieran aplicado los niveles del factor principal.

Como en los ejemplos anteriores, el que una unidad experimental pertenezca a un rebaño o a otro no es aleatorio, sino que es el experimentador el que lo determina con arreglo a criterios ajenos al problema; normalmente esta segunda fuente de variación también es debida a factores ambientales o a criterios de clasificación. Y a diferencia de los ejemplos anteriores, los tratamientos se aplican aleatoriamente a los niveles del segundo factor y no dentro de cada nivel del segundo factor. En este

ejemplo concreto, se han tenido que elegir aleatoriamente seis rebaños para aplicarle aleatoriamente uno de los tres tratamientos a cada dos. Por lo tanto, el que un rebaño reciba uno u otro tratamiento es aleatorio.

Tal como se ha hecho con los ejemplos anteriores, la relación existe entre estos valores puede expresarse por medio de desviaciones. Si hacemos referencia al ejemplo anterior (dos factores con repetición) aquí las desviaciones son parecidas pues también hay dos factores. De manera que la desviación total de una unidad experimental (valor de dicha unidad menos la media general) es igual a la desviación debida al tratamiento en que está dicho individuo más la desviación debida al rebaño en que está, más la variación individual. En este caso, el primer factor o tratamientos es igual que en los ejemplos anteriores, pero el segundo factor difiere de los ejemplos anteriores, pues los niveles de este segundo factor son privativos de los niveles del factor principal, por lo que las desviaciones de este segundo factor debe ser referidas a la media del grupo o nivel del factor principal en el que se encuentran; de la misma manera que la desviación individual es con respecto al grupo en que se encuentran los individuos.

Resumiendo, cada individuo pertenece a un grupo o rebaño, cada rebaño pertenece a un grupo o tratamiento, por lo que la desviación total será igual a la *i-ésima* desviación del grupo  $T_i$  más la *j-ésima* desviación de grupo  $R_j$  dentro del grupo  $T_i$ ,  $R_{j(i)}$  más la desviación individual, es decir

$$(X_{ijk} - \bar{X}_{...}) = (\bar{X}_{i..} - \bar{X}_{...}) + (\bar{X}_{ij.} - \bar{X}_{i..}) + (X_{ijk} - \bar{X}_{ij.})$$

Pasando la media general del termino de la derecha al de la izquierda, se obtiene

$$X_{ijk} = \bar{X}_{...} + (\bar{X}_{i..} - \bar{X}_{...}) + (\bar{X}_{ij.} - \bar{X}_{i..}) + (X_{ijk} - \bar{X}_{ij.})$$

Lo que quiere decir que la medida de una unidad experimental es igual a la suma de cuatro sumandos, el primero es la estima de la *media general*,  $\mu$ , el segundo es la estima del efecto del *tratamiento*,  $\tau_i$ , el tercero es la estima del efecto de la *raza j* dentro del tratamiento *i*,  $\rho_{j(i)}$ , y el cuarto es la estima del *error experimental*,  $\varepsilon_{ij}$ . Por lo que en este caso el modelo lineal es

$$X_{ijk} = \bar{X}_{...} + (\bar{X}_{i..} - \bar{X}_{...}) + (\bar{X}_{ij.} - \bar{X}_{i..}) + (X_{ijk} - \bar{X}_{ij.})$$

$$X_{ijk} = \mu + T_i + R_{j(i)} + \varepsilon_{ijk}$$

Recuérdese que, en este ejemplo, el primer factor se trata de tres *tratamientos* que pueden ser tres dietas o tres antibióticos o tres tóxicos o tres métodos de purificación, etc., es decir, un factor con tres únicos niveles del cual se quiere saber si existen diferencias entre dichos niveles y en caso positivo, saber cuál o cuales es el mejor. Con respecto al segundo factor, *rebaño*, no ocurre lo mismo sino que este segundo factor lo ha introducido el experimentador obligado por criterios ambientales o de clasificación. El experimentador no tiene interés en saber si existen diferencias entre los niveles de este segundo factor, se introduce en el modelo para explicar mejor la variabilidad observada. Como se ve, este ejemplo es semejante al Enunciado 2 del

epígrafe *Factores de efectos fijos y factores de efectos aleatorios*, con la diferencia de que, aquí, un tratamiento se aplica en unos rebaños y otro tratamiento se aplicará a otros rebaños, aunque los rebaños serán semejantes; por tanto el modelo que lo explica es mixto, es decir, los tratamientos son fijos y los rebaños son aleatorios.

Como  $T_i$  es de efectos fijos y  $R_{j(i)}$  es de efectos aleatorios, se tiene que

$$\begin{aligned}\sum_i T_i &= 0 \\ R_{j(i)} &\sim N(0, \sigma_R^2) \\ e_{ijk} &\sim N(0, \sigma_R^2)\end{aligned}$$

por lo que se tiene dos varianzas: la del factor aleatorio, rebaño, y la del error que estima la de la población.

Como  $R_{j(i)}$  es aleatorio, significa que la sumas de sus efectos no tiene porque valer cero, y estos son función de la varianza de este factor, por lo que la variabilidad de los tratamientos se estima a través de la variabilidad de los rebaños y ésta a través de la variabilidad de los individuos. Por lo que la variabilidad de rebaño incluye, además de la debida a dicho factor, la debida al error. Y la variabilidad de los tratamientos incluye, además de la debida a dicho factor la variabilidad de los rebaños y la del error.

Por tanto, en este caso, hay que comparar la variabilidad de los tratamientos con la de los rebaños, y la de los rebaños con la del error. Si se hace así se comprueba que la variabilidad de los tratamientos es mucho mayor que la de los rebaños que es prácticamente nula, por lo que se está más seguros de afirmar la existencia de diferencias entre tratamientos, cuando en el primer ejemplo, sin tener en cuenta los rebaños, no estábamos seguro de esta afirmación. Para los rebaños dentro de tratamientos se puede afirmar que no hay diferencia, pues la variabilidad es prácticamente nula, siendo la variabilidad individual mucho mayor.

El modelo podría haber sido aleatorio, es decir, tratamientos y rebaños de efectos aleatorios, semejante al Enunciado 3 del epígrafe *Factores de efectos fijos y factores de efectos aleatorios*, con la diferencia de que aquí, cada uno de los tres macho (los tres tratamientos) tiene descendencia en dos rebaños, que no son los mismos para los tres machos, con arreglo a la tabla de datos de este ejemplo.

Como  $T_i$  y  $R_{j(i)}$  son de efectos aleatorios, se tiene que

$$\begin{aligned}T_i &\sim N(0, \sigma_T^2) \\ R_{j(i)} &\sim N(0, \sigma_R^2) \\ e_{ijk} &\sim N(0, \sigma^2)\end{aligned}$$

por lo que se tiene tres varianzas: las de los dos factores aleatorios, tratamientos y rebaños, y la del error que estima la de la población.

Como  $R_{j(i)}$  es aleatorio, la variabilidad de los tratamientos se estima a través de la variabilidad de los rebaños y ésta a través de la variabilidad de los individuos, por lo que se llega a la mismas conclusiones que con el modelo mixto, estas son, que la variabilidad de los tratamientos incluye la variabilidad de los rebaño y la del error, y que la variabilidad del rebaño solo incluye, además de la suya, la del error.

Por tanto, en este caso, como en el anterior, hay que comparar la variabilidad de los tratamientos con la de los rebaños, y la de los rebaños con la del error. Si se hace así se comprueba que la variabilidad de los tratamientos es mucho mayor que la de los rebaños que es prácticamente nula, por lo que se está más seguros de afirmar la existencia de diferencias entre tratamientos, cuando en el primer ejemplo, sin tener en cuenta los rebaños, no estábamos seguro de esta afirmación. Para los rebaños dentro de tratamientos se puede afirmar que no hay diferencia, pues la variabilidad es prácticamente nula, siendo la variabilidad individual mucho mayor.

Aunque poco corriente en los modelos jerárquicos, puede ocurrir que el modelo hubiera sido fijo, es decir, tanto el factor  $T_i$  como el factor  $R_{j(i)}$  sean de efectos fijos, semejante al Enunciado 1 del epígrafe *Factores de efectos fijos y factores de efectos aleatorios*, pero aquí, los tres tratamientos puede ser tres ciudades o tres comarcas, y las dos razas pueden ser dos rebaños (dentro de cada comarca, por tanto un total de seis rebaños) o dos hospitales o dos planta de hospital o dos centros de distribución (dentro de cada ciudad) siendo las tres ciudades o tres comarcas las únicas que nos interesan y los dos rebaños o dos hospitales los únicos que nos interesan.

Como tanto  $T_i$  como  $R_{j(i)}$  son de efectos fijos, se tiene que

$$\begin{aligned} \sum_i T_i &= 0 \\ \sum_{ij} R_{j(i)} &= 0 \\ e_{ijk} &\sim N(0, \sigma^2) \end{aligned}$$

por lo que se tiene solo una varianza, la del error, que estima la de la población.

Al no haber más varianza que la del error, la variabilidad de los dos factores solo incluirían, además de la propia, la del error. Por lo que en este caso, hay que comparar la variabilidad de los tratamientos y de los rebaños con la del error. Si se hace así se comprueba que la variabilidad de los tratamientos no es mucho mayor que la variabilidad individual por lo que hubiéramos tenido las mismas dudas que en el primer ejemplo (con solo los tratamientos) de que la variabilidad de los tratamientos no era mayor que la de las casillas o error, por lo que no se podría afirmar que existen diferencias entre los tratamientos. Para los rebaños dentro de tratamientos, como en los casos anteriores, se puede afirmar que no hay diferencia, pues la variabilidad es prácticamente nula, siendo la variabilidad individual mucho mayor.

Recuérdese que se está tratando de hacer un análisis intuitivo, el análisis estadístico se hará más adelante.

## Diseños Jerárquicos o con subgrupos.-

El motivo por lo que se realizan diseños jerárquicos es porque en algunas situaciones experimentales es insuficiente el realizar un solo grupo de réplicas de un tratamiento o nivel del factor, por lo que se pueden realizar varios subgrupos de observaciones dentro del grupo al que se le aplica el tratamiento. Tales observaciones se hacen en subgrupos o unidades de muestreo. Las diferencias entre subgrupos dentro de un nivel del factor son diferencias de observación o de unidad de medida más que diferencias de unidad experimental. Por ejemplo, considérese el experimento en el que se quiere probar cuatro insecticidas para el ganado, para ello se toman doce animales y aleatoriamente se separan en cuatro grupos entre los cuales se determinará aleatoriamente cuál es el *tratamiento/insecticida* que se le aplica. La aplicación del insecticida se hace rociando cada individuo con el insecticida que le ha correspondido y midiéndole, después de un tiempo, el número de parásitos que se encuentran en diferentes (cuatro) zonas del cuerpo.

En un experimento con un solo animal para cada insecticida, nunca se estaría seguros de que las diferencias observadas entre animales se debieran a la diferente composición del insecticida. Posiblemente, diferencias fisiológicas y otras accidentales entre los individuos podrían dar como resultado diferencias de la acción insecticida entre los cuatro productos, las cuales podrían presentarse, incluso si hubiéramos usado el mismo insecticida. Se tiene, por tanto, dos efectos: el *insecticida* y los *animales*; la única manera de separarlos es tener dos o más animales para cada insecticida.

De las dos fuentes de variación del experimento descrito en el párrafo anterior

### ***Variación entre insecticidas y***

### ***Variación entre animales dentro de insecticida.***

Esta última fuente de variación, contiene la *variación debida a la variabilidad individual* (de las unidades experimentales, que en este caso son las zonas de medida) dentro de los animales, es decir, el ***error experimental*** y la *variación debida a la variabilidad entre los animales*, es decir, el ***error de muestreo***. La variabilidad entre los animales dentro de tratamiento no tiene porqué ser mayor que la variabilidad dentro de los animales, pero lo corriente es que sí lo sea y como consecuencia de ello se introduce en el experimento.

Si no encontramos diferencia entre los animales dentro de insecticida (mayor que la diferencia dentro de animal) podremos atribuir las diferencias entre los insecticidas a sus composiciones. Pero, incluso si encontrásemos que existe una cantidad significativa de varianza entre los animales dentro de insecticida, se puede contrastar si la variación entre los insecticidas es mayor de lo esperado por el azar sobre la base de la variación observada entre los individuos dentro de insecticida. Si el cuadrado medio entre insecticida es significativamente mayor que el cuadrado medio de entre animales dentro de insecticida podemos concluir que las diferencias entre insecticidas son debidas a causas no aleatorias.

Un diseño de este tipo es un diseño jerárquico o encajado o anidado, puesto que el criterio de clasificación subordinado es aplicable dentro de las categorías del criterio superior. Así, en nuestro ejemplo, los animales están subordinados al insecticida. Un aspecto importante es que los grupos que representan un nivel subordinado de clasificación, corrientemente, se escogen aleatoriamente. Por tanto, casi siempre, el factor subordinado de un diseño jerárquico es de efectos aleatorios, mientras que el superior puede ser tanto de efectos fijos como aleatorios. Si se tuviera un diseño jerárquico en el que todos los factores son aplicables dentro de los niveles de los correspondientes factores superiores (menos el factor principal) y todos fueran de efectos aleatorios, tendríamos un diseño encajado puro.

Existen dos clases de aplicación de diseños jerárquicos. La primera es la que ya hemos visto con el ejemplo de los insecticidas. Esta clase de aplicación sirve para asegurar la magnitud mínima del error en los diversos pasos de un experimento o un proceso industrial. De manera que en el ejemplo de los insecticidas, se desea averiguar si existe una variación significativa tanto entre los animales como dentro de los animales, y se necesita conocer la magnitud de esta variabilidad con objeto de poder contrastar la magnitud de la componente añadida debida a los *tratamientos/insecticidas*. El diseño jerárquico no está limitado a los dos niveles descritos hasta ahora, dado que se pueden subdividir los grupos a su vez en subgrupos, y así sucesivamente, continuando mientras se puedan seguir subdividiendo aleatoriamente los subgrupos. De esta manera se podría diseñar un experimento para comprobar los efectos de cinco productos químicos sobre la cantidad de pigmentación en la piel de los mamíferos. Los *cinco productos* y un *control* (seis grupos en total) son los niveles del factor superior y, claramente, son tratamientos de efectos fijos. Para cada producto se podrían escoger *cinco ratas* seleccionadas aleatoriamente. Dichas ratas proporcionarían una medida de la *varianza de las ratas dentro de una clase de producto* químico o de rata dentro de producto. De cada rata podríamos tomar, aleatoriamente, tres muestras de piel de su zona ventral. Esto significa una nueva fuente de variación subordinada (muestra de *piel dentro de rata*). Cada muestra de piel se macera y divide en dos lotes, que se hidrolizan por separado. Este nivel está formado por las *partes hidrolizadas dentro de las muestras de piel*. Finalmente, la cantidad de pigmentación podría ser leída como la densidad óptica, pudiéndose hacer dos lecturas repetidas de cada parte hidrolizada. El *error experimental* sería la *varianza de las lecturas repetidas por parte hidrolizada*, pero también se tendría valores de la *varianza entre las partes hidrolizadas dentro de una muestra de piel, varianza entre las muestras de piel dentro de una rata y varianza entre ratas dentro de un producto químico*. Estos valores de la varianza son importantes en el diseño de experimentos análogos porque indican en qué nivel del experimento se deberá concentrar la mayor parte del esfuerzo. El aspecto más variable del experimento necesitará mayor número de repeticiones o un control experimental más estricto. De manera, que si se encuentra que la mayor proporción de variación es para las dos partes hidrolizadas es porque el método utilizado para hidrolizar no está muy perfeccionado y debería, por tanto, ser mejorado. Si esto no fuera posible, se debería de dividir cada muestra de piel macerada en más lotes, a fin de disponer de más partes hidrolizadas.

La segunda clase de aplicación de los diseños jerárquicos es para los jerárquicos puros. Estos casos son corrientes en muchas ramas de la ciencia donde se desea conocer la magnitud de las componentes de la varianza atribuible a los diversos niveles de variación. Por ejemplo, en genética animal, cada uno de cierto número de

machos se aparean con varias hembras y varios miembros de la progenie de cada hembra se miden para obtener los datos. Los individuos así medidos forman una población de familias de medios hermanos y de hermanos carnales. Aquí, las fuentes de variación son: *entre machos, entre hembras dentro de machos y dentro de hembras o progenies*, el error. Y se desea conocer las componentes debidas a los machos, a las hembras y a los descendientes de cada hembra. Estas componentes nos las dan las ECM.

### Ejecución del diseño.-

Ya se ha expuesto anteriormente que si se tiene  $t$  tratamientos hay que elegir aleatoriamente los componentes de cada uno de los  $t$  subgrupos y después aleatorizar cuál de los  $t$  tratamientos se le aplica a cada subgrupo.

### Modelo lineal.-

Como ya se ha dicho, puesto que la variación existente entre unidades experimentales es atribuible a

El error experimental.

El error de muestreo o la varianza de los subgrupos dentro de tratamientos y

A los tratamientos,

El modelo lineal que explica esta variación es

$$X_{ijk} = \mu + T_i + R_{j(i)} + e_{k(ij)}$$

donde  $\mu$  es la media general,  $T_i$  es el efecto del  $i$ -ésimo nivel del factor superior o *principal* o el  $i$ -ésimo tratamiento,  $R_{j(i)}$  es el efecto del  $j$ -ésimo nivel del factor subordinado dentro del  $i$ -ésimo nivel del factor  $T$  y  $e_{k(ij)}$  es el efecto del individuo  $k$  del nivel  $j$  de  $R$  y del nivel  $i$  de  $T$ .

### Descomposición de la varianza en un modelo jerárquico.-

Siguiendo con el mismo criterio de los modelos anteriores, definimos los efectos como

$$T_i = \bar{X}_{i..} - \bar{X}_{...}$$

$$R_{j(i)} = \bar{X}_{ij.} - \bar{X}_{i..}$$

y establecemos la identidad

$$X_{ijk} - \bar{X}_{...} = (\bar{X}_{i..} - \bar{X}_{...}) + (\bar{X}_{ij.} - \bar{X}_{i..}) + (X_{ijk} - \bar{X}_{ij.})$$

elevando al cuadrado esta expresión y teniendo en cuenta que los productos cruzados se anulan,

$$\sum_{ijk} (X_{ijk} - \bar{X}_{...})^2 = n_{ij} \sum_i n_i (\bar{X}_{i..} - \bar{X}_{...})^2 + n_{ij} \sum_{ij} (\bar{X}_{ij.} - \bar{X}_{i..})^2 + \sum_{ijk} (X_{ijk} - \bar{X}_{ij.})^2$$

donde el primer sumando nos mide la variación total entre los niveles de  $T$ . El segundo sumando nos mide la variación total entre los niveles de  $R$  dentro de cada nivel de  $T$ . Y el último sumando es la suma de cuadrados de los errores.

Con lo que se tiene el siguiente cuadro resumen

$SC_{Trata}$	$n_{ij} \sum_i n_i (\bar{X}_{i..} - \bar{X}_{...})^2$	$=$	$\sum_i \frac{X_{i..}^2}{n_{i..}} - \frac{X_{...}^2}{N}$
$SC_{Reb(T)}$	$n_{ij} \sum_{ij} (\bar{X}_{ij.} - \bar{X}_{i..})^2$	$=$	$\sum_{ij} \frac{X_{ij.}^2}{n_{ij.}} - \sum_i \frac{X_{i..}^2}{n_{i..}}$
$SC_{Error}$	$\sum_{ijk} (X_{ijk} - \bar{X}_{ij.})^2$	$=$	$\sum_{ijk} X_{ijk}^2 - \sum_{ij} \frac{X_{ij.}^2}{n_{ij.}}$
$SC_{total}$	$\sum_{ijk} (X_{ijk} - \bar{X}_{...})^2$	$=$	$\sum_{ijk} X_{ijk}^2 - \frac{X_{...}^2}{N}$

En la composición de estas expresiones se ve, como en los modelos anteriores, que hay una lógica, cada sumando elevado al cuadrado está dividido por el número de elementos que han formado esta suma, por ejemplo, si han intervenido los  $n_i$  números que constituyen los individuos de los  $i$ -ésimos tratamientos, se dividirá por  $n_i$ ; si la suma es de  $n_{ij}$  individuos, del  $ij$ -ésimo rebaño, hay que dividir por  $n_{ij}$ ; y si la suma es de  $N$  individuos, hay que dividir por  $N$ , etc. Esta misma lógica servirá para saber los grados de libertad que se verán en el siguiente epígrafe.

### Descomposición de los grados de libertad en un modelo jerárquico.-

Siguiendo con los criterios establecidos en el modelo de una vía, para el factor principal los grados de libertad son el número de niveles de dicho factor menos uno. Para el factor o factores jerárquicos, los grados de libertad son el número de niveles totales de dicho factor, menos el número de niveles del factor superior. Y para el error, los grados de libertad son el número total de datos que estamos analizando, menos el número total de niveles del último factor jerárquico.

Si se sigue denominando como  $t$  los niveles del primer factor (el  $T$  en este caso) y como  $r$  los niveles del factor subordinado dentro del principal, se puede denominar como  $R$  el número total de niveles del factor subordinado, de manera que se tendrá el siguiente cuadro resumen



<i>FV</i>	<i>gl</i>	<i>SC</i>
<i>Trata</i>	$t-1$	$\sum_i \frac{X_{i..}^2}{n_{i..}} - \frac{X_{...}^2}{N}$
<i>Rebaño(T)</i>	$R-t$	$\sum_{ij} \frac{X_{ij.}^2}{n_{ij.}} - \sum_i \frac{X_{i..}^2}{n_{i..}}$
<i>Error</i>	$N-R$	$\sum_{ijk} X_{ijk}^2 - \sum_{ij} \frac{X_{ij.}^2}{n_{ij.}}$
<i>total</i>	$N-1$	$\sum_{ijk} X_{ijk}^2 - \frac{X_{...}^2}{N}$

### Parámetros estimados en un modelo jerárquico.-

Recuérdese que cuando se realizaron pruebas de hipótesis en los modelos de una vía y modelos factoriales sin repetición, las *F* de prueba son siempre las mismas, sea cual sea el tipo o tipos de factores que se tengan. Y recuerdes, así mismo, que en el caso del modelo factorial con repetición, las esperanzas matemáticas de los cuadrados medios de los diferentes factores son diferentes según el tipo de factor. Repásese lo que se dijo en los epígrafes *Modelo de efectos fijos*, *Modelo de efectos aleatorios* y *Modelo mixto*.

Lo habitual en modelos jerárquicos es que sean modelos mixtos, el factor principal fijo y el factor jerárquico aleatorio, o modelos aleatorios. El factor jerárquico, normalmente, es aleatorio, pero también se puede dar el caso de que sea fijo, y si el factor jerárquico es fijo, necesariamente, es fijo el factor principal, por lo que se tendría un modelo fijo.

Supóngase, ahora, que se tiene un *Modelo Aleatorio* (ver epígrafe), es decir, tratamientos y rebaños de efectos aleatorios, semejante al Enunciado 3 del epígrafe *Factores de efectos fijos y factores de efectos aleatorios*, con la diferencia de que aquí, cada uno de los tres macho (los tres tratamientos) tiene descendencia en dos rebaños, que no son los mismos para los tres machos, con arreglo a la tabla de datos de este ejemplo.

Por lo que se tienen tres varianzas, estas son

$$T_i \sim N(0, \sigma_T^2)$$

$$R_{j(i)} \sim N(0, \sigma_R^2)$$

$$e_{ijk} \sim N(0, \sigma^2)$$

En el modelo aleatorio, por tanto, la estima de los efectos del factor principal se realiza a través de una muestra aleatoria del universo de niveles del factor jerárquico.

Por lo que al estimar la variabilidad del factor principal (*machos*), se estará haciendo a través de la variabilidad del factor jerárquico y de la variabilidad individual (error), esto hace que las variabilidades de machos, incluyan, además de la debida a dicho factor, la variabilidad del factor rebaño y la del error. Mientras que la estima de los efectos del factor jerarquizado se hará a través de la variabilidad individual, por lo que incluirá, además de la varianza de dicho factor, la varianza del error.

Calculando, por tanto, los valores que estiman las sumas de cuadrados de manera análoga a como se hizo con el modelo de una sola vía, y al dividir las estimas de las *SC* por los *gl* se obtienen las de los *CM*, que son

$$ECM_T = \sigma^2 + n\sigma_R^2 + n_{ij}\sigma_T^2$$

$$ECM_{R(T)} = \sigma^2 + n\sigma_R^2$$

$$ECM_{Error} = \sigma^2$$

Quedando la tabla resumen de la siguiente manera  
Modelo Aleatorio

<i>FV</i>	<i>gl</i>	<i>SC</i>	<i>ECM</i>
<i>Trata</i>	<i>t-1</i>	$\sum_i \frac{X_{i..}^2}{n_{i..}} - \frac{X_{...}^2}{N}$	$\sigma^2 + n\sigma_R^2 + n_{ij}\sigma_T^2$
<i>Rebaño(T)</i>	<i>R-t</i>	$\sum_{ij} \frac{X_{ij.}^2}{n_{ij.}} - \sum_i \frac{X_{i..}^2}{n_{i..}}$	$\sigma^2 + n\sigma_R^2$
<i>Error</i>	<i>N-R</i>	$\sum_{ijk} X_{ijk}^2 - \sum_{ij} \frac{X_{ij.}^2}{n_{ij.}}$	$\sigma^2$
<i>Total</i>	<i>N-1</i>	$\sum_{ijk} X_{ijk}^2 - \frac{X_{...}^2}{N}$	

Supóngase, ahora, que se tiene un *Modelo Mixto* (ver epígrafe) semejante al Enunciado 2 del epígrafe *Factores de efectos fijos y factores de efectos aleatorios*, esto es, los tres niveles del factor principal (tres *tratamientos*) son el universo de niveles de este factor, no hay más tratamientos o no interesan otros, si los hubiera, y los dos niveles del factor jerárquico (dos *rebaños* por tratamiento) son una muestra aleatoria de los muchos niveles existentes. Como se observa, la diferencia con respecto a un factorial, es que, aquí, un tratamiento concreto se aplica en unos rebaños y otro tratamiento se aplicará a otros rebaños, aunque los rebaños serán semejantes.

Por lo que se tienen dos varianzas, estas son

$$\sum_i T_i = 0$$

$$R_{j(i)} \sim N(0, \sigma_R^2)$$

$$e_{ijk} \sim N(0, \sigma^2)$$

En el modelo mixto, por tanto, la estima de los efectos del factor principal se realiza a través de una muestra aleatoria del universo de niveles del factor jerárquico.

Por lo que al estimar la variabilidad del factor principal (*tratamientos*), se estará haciendo a través de la variabilidad del factor jerárquico y de la variabilidad individual (error), esto hace que la variabilidad de los tratamientos, incluyan, además de la debida a dicho factor, la variabilidad del factor rebaño y la del error. Mientras que la estima de los efectos del factor jerarquizado se hará a través de la variabilidad individual, por lo que incluirá, además de la varianza de dicho factor, la varianza del error.

Calculando, por tanto, los valores que estiman las sumas de cuadrados de manera análoga a como se hizo con el modelo de una sola vía, y al dividir las estimas de las *SC* por los *gl* se obtienen las de los *CM*, que son

$$ECM_T = \sigma^2 + n \sigma_R^2 + n_{ij} \cdot \sigma_T^2$$

$$ECM_{R(T)} = \sigma^2 + n \sigma_R^2$$

$$ECM_{Error} = \sigma^2$$

Se ve que estas estimas son las mismas que las del modelo aleatorio, como consecuencia de, como dijimos anteriormente, cuando se trate del factor en cuestión, aunque este sea de efectos fijos, en sus componentes se pondrá la debida a sus efectos, como una varianza.

Quedando la tabla resumen de la siguiente manera

Modelo Mixto

<i>FV</i>	<i>gl</i>	<i>SC</i>	<i>ECM</i>
<i>Trata</i>	<i>t-1</i>	$\sum_i \frac{X_{i..}^2}{n_{i..}} - \frac{X_{...}^2}{N}$	$\sigma^2 + n\sigma_R^2 + n_{ij} \cdot n\sigma_T^2$
<i>Rebaño(T)</i>	<i>R-t</i>	$\sum_{ij} \frac{X_{ij.}^2}{n_{ij.}} - \sum_i \frac{X_{i..}^2}{n_{i..}}$	$\sigma^2 + n\sigma_R^2$
<i>Error</i>	<i>N-R</i>	$\sum_{ijk} X_{ijk}^2 - \sum_{ij} \frac{X_{ij.}^2}{n_{ij.}}$	$\sigma^2$
<i>Total</i>	<i>N-1</i>	$\sum_{ijk} X_{ijk}^2 - \frac{X_{...}^2}{N}$	

Supóngase que, aunque poco frecuente en modelos jerárquicos, se trata de un *Modelo fijo* (ver epígrafe), es decir, tanto el factor  $T_i$  como el factor  $R_{j(i)}$  son de efectos fijos, semejante al Enunciado 1 del epígrafe *Factores de efectos fijos y factores de efectos aleatorios*, pero aquí, los tres tratamientos puede ser tres ciudades o tres comarcas, y las dos razas pueden ser dos rebaños (dentro de cada comarca, por tanto un total de seis rebaños) o dos hospitales o dos planta de hospital o dos centros de distribución (dentro de cada ciudad) siendo las tres ciudades o tres comarcas las únicas que nos interesan y los dos rebaños o dos hospitales los únicos que nos interesan, y se quiere saber si son iguales o diferentes, tanto unos como los otros.

En este modelo fijo, el factor principal, los *tratamientos*, se estiman a través de todo el universo de valores de rebaños, y estos se estiman a partir de la variabilidad de las casillas o error, por lo que la única varianza que se tiene es la de la población que, como en todos los demás modelos, es estimada por el error de manera que

Como tanto  $T_i$  como  $R_{j(i)}$  son de efectos fijos, se tiene que

$$\begin{aligned} \sum_i T_i &= 0 \\ \sum_{ij} R_{j(i)} &= 0 \\ e_{ijk} &\sim N(0, \sigma^2) \end{aligned}$$

por lo que se tiene solo una varianza, la del error, que estima la de la población.

Por lo que al estimar la variabilidad de las dos fuentes de variación (*tratamiento y rebaño*), se estará haciendo a través de la variabilidad del error, esto hace que estas variabilidades incluyan, además de la debida a dichos factores, la variabilidad del error.

Calculando, por tanto, los valores que estiman las sumas de cuadrados de manera análoga a como se hizo con el modelo de una sola vía, y al dividir las estimas de las *SC* por los *gl* se obtienen las de los *CM*, que son

$$ECM_T = \sigma^2 + n_{ij} \sigma_T^2$$

$$ECM_R = \sigma^2 + n \sigma_R^2$$

$$ECM_{Error} = \sigma^2$$

Quedando la tabla resumen de la siguiente manera

Modelo Fijo			
FV	gl	SC	ECM
<i>Trata</i>	<i>t-1</i>	$\sum_i \frac{X_{i..}^2}{n_{i..}} - \frac{X_{...}^2}{N}$	$\sigma^2 + n_{ij} \cdot n \sigma_T^2$
<i>Rebaño(T)</i>	<i>R-t</i>	$\sum_{ij} \frac{X_{ij.}^2}{n_{ij.}} - \sum_i \frac{X_{i..}^2}{n_{i..}}$	$\sigma^2 + n \sigma_R^2$
<i>Error</i>	<i>N-R</i>	$\sum_{ijk} X_{ijk}^2 - \sum_{ij} \frac{X_{ij.}^2}{n_{ij.}}$	$\sigma^2$
<i>Total</i>	<i>N-1</i>	$\sum_{ijk} X_{ijk}^2 - \frac{X_{...}^2}{N}$	

### Pruebas de hipótesis en un modelo jerárquico.-

Con arreglo al modelo lineal planteado, se pueden plantear dos posibles pruebas de hipótesis, estas son las de contrastar si los efectos debidos a los diferentes niveles de los dos factores son o no significativos. Como es un modelo jerárquico no es posible estimar la interacción, esta no se puede contrastar. Por tanto las pruebas pueden ser

Para el factor principal

$$H_0 : \sigma_T^2 = 0$$

$$H_1 : \sigma_T^2 > 0$$

Para el factor jerárquico

$$H_0 : \sigma_R^2 = 0$$

$$H_1 : \sigma_R^2 > 0$$

Para realizar la prueba *F*, que determine la posible significación de un factor o fuente de variación (que la varianza producida por el factor es mayor de cero) se toma como numerador el *CM* de dicho factor y como denominador el *CM* de la fuente de

variación que en su  $ECM$  presenta las mismas componentes de la varianza del numerador, afectadas de iguales coeficientes, menos la componente que se está contrastando si es igual a cero.

El  $CM$  que se usa como denominador para la prueba  $F$  se suele denominar *término de error* para el factor que se está probando; que no tiene por qué ser el término llamado *error* que es realmente el que mide la variabilidad residual.

Por tanto, las posibles pruebas  $F$  que se pueden realizar en modelos jerárquicos, son diferentes según el tipo de factores que se tengan, pues según el tipo de factor, sus cuadrados medios estimarán diferentes parámetros y por lo tanto, tendrán como término de error diferentes cuadrados medios. Aunque, como se ha dicho anteriormente, lo corriente es que el factor jerárquico (o factores jerárquicos) sea aleatorio; es el factor principal el que puede ser fijo o aleatorio, lo cual da lo mismo a efectos de pruebas de hipótesis, por lo que se van a ver ambos modelos juntos.

Se verán juntos el Modelo Mixto y el Aleatorio.

Se seguirá simbolizando el factor principal como  $T$ .

En el caso de la prueba del primer factor, si  $\sigma_T^2 = 0$

$$ECM_{Trata} = \sigma^2 + n\sigma_R^2 + n_{ij} \cdot n\sigma_T^2 = \sigma^2 + n\sigma_R^2 + n_{ij} \cdot n0^2 = \sigma^2 + n\sigma_R^2 = ECM_R$$

Es decir, el cuadrado medio del factor principal estima lo mismo que el cuadrado medio del factor jerárquico, luego, si se dividen ambos cuadrados medios, se espera un valor de  $F$  no significativo, esto es, estadísticamente igual a uno.

Y si  $\sigma_T^2 > 0$

$$ECM_{Trata} = \sigma^2 + n\sigma_R^2 + n_{ij} \cdot n\sigma_T^2 > \sigma^2 + n\sigma_R^2 = ECM_R$$

Es decir, el cuadrado medio de los tratamientos estima una componente más (la debida a los tratamientos) que el cuadrado medio del factor jerárquico, luego, si se dividen ambos cuadrados medios, se espera un valor de  $F$  significativo, esto es, estadísticamente superior a uno.

Por lo que el estadístico de prueba para el factor principal será

$$F_o = \frac{CM_{Tratamiento}}{CM_{R(T)}}$$

En el caso de la prueba del segundo factor, si  $\sigma_R^2 = 0$

$$ECM_{Rebaño} = \sigma^2 + n\sigma_R^2 = \sigma^2 + n0^2 = ECM_{Error}$$

Es decir, el cuadrado medio del factor jerárquico estima lo mismo que el cuadrado medio del error, luego, si se dividen ambos cuadrados medios, se espera un valor de  $F$  no significativo, esto es, estadísticamente igual a uno.

Y si  $\sigma^2_R > 0$

$$ECM_{Rebaño} = \sigma^2 + n \sigma^2_R > \sigma^2 = ECM_{Error}$$

Es decir, el cuadrado medio de los rebaños estima una componente más (la debida a los rebaños) que el cuadrado medio del error, luego, si se dividen ambos cuadrados medios, se espera un valor de  $F$  significativo, esto es, estadísticamente superior a uno.

Por lo que el estadístico de prueba para el factor jerárquico será

$$F_o = \frac{CM_{Rebaño(Tr)}}{CM_{Error}}$$

Quedando el análisis de esta manera

#### Modelo Mixto y Aleatorio

<i>FV</i>	<i>gl</i>	<i>SC</i>	<i>CM</i>	<i>ECM</i>	<i>F<sub>o</sub></i>
<i>Trata</i>	<i>t-1</i>	$\sum_i \frac{X_{i..}^2}{n_{i..}} - \frac{X_{...}^2}{N}$	$\frac{SCTr}{t-1}$	$\sigma^2 + n \sigma^2_R + n_{ij} \cdot n \sigma^2_T$	$\frac{CM_{Trata}}{CM_{Rebaño}}$
<i>Rebaño(T)</i>	<i>R-t</i>	$\sum_{ij} \frac{X_{ij.}^2}{n_{ij.}} - \sum_i \frac{X_{i..}^2}{n_{i..}}$	$\frac{SCRe}{R-t}$	$\sigma^2 + n \sigma^2_R$	$\frac{CM_{Reba}}{CM_{Error}}$
<i>Error</i>	<i>N-R</i>	$\sum_{ijk} X_{ijk}^2 - \sum_{ij} \frac{X_{ij.}^2}{n_{ij.}}$	$\frac{SCEr}{N-R}$	$\sigma^2$	
<i>total</i>	<i>N-1</i>	$\sum_{ijk} X_{ijk}^2 - \frac{X_{...}^2}{N}$			

Por ultimo, el poco corriente Modelo Fijo.

En el caso de la prueba del factor principal, si  $\sigma^2_T = 0$

$$ECM_{Tratamiento} = \sigma^2 + n_{ij} \cdot n \sigma^2_T = \sigma^2 + n_{ij} \cdot n \cdot 0 = \sigma^2 = ECM_{Error}$$

Es decir, el cuadrado medio del factor principal estima lo mismo que el cuadrado medio del error, luego, si se dividen ambos cuadrados medios, se espera un valor de  $F$  no significativo, esto es, estadísticamente igual a uno.

Y si  $\sigma^2_T > 0$

$$ECM_{Tratamiento} = \sigma^2 + n_{ij} \cdot n \sigma^2_T > \sigma^2 = ECM_{Error}$$

Es decir, el cuadrado medio de los tratamientos estima una componente más (la debida a los tratamientos) que el cuadrado medio del error, luego, si se dividen ambos cuadrados medios, se espera un valor de  $F$  significativo, esto es, estadísticamente superior a uno.

Por lo que el estadístico de prueba para el factor principal será

$$F_o = \frac{CM_{Tratamiento}}{CM_{Error}}$$

En el caso de la prueba del segundo factor, si  $\sigma_R^2 = 0$

$$ECM_{Rebaño} = \sigma^2 + n \sigma_R^2 = \sigma^2 + n 0^2 = \sigma^2 = ECM_{Error}$$

Es decir, el cuadrado medio del factor jerárquico estima lo mismo que el cuadrado medio del error, luego, si se dividen ambos cuadrados medios, se espera un valor de  $F$  no significativo, esto es, estadísticamente igual a uno.

Y si  $\sigma_R^2 > 0$

$$ECM_{Rebaño} = \sigma^2 + n \sigma_R^2 > \sigma^2 = ECM_{Error}$$

Es decir, el cuadrado medio de las razas estima una componente más (la debida a las razas) que el cuadrado medio del error, luego, si se dividen ambos cuadrados medios, se espera un valor de  $F$  significativo, esto es, estadísticamente superior a uno.

Por lo que el estadístico de prueba para el factor jerárquico será

$$F_o = \frac{CM_{Rebaño(Tr)}}{CM_{Error}}$$



Quedando el análisis de esta manera

Modelo Fijo

FV	gl	SC	CM	ECM	F <sub>o</sub>
Tratamie	t-1	$\sum_i \frac{X_{i..}^2}{n_{i..}} - \frac{X_{...}^2}{N}$	$\frac{SC_{Tr}}{t-1}$	$\sigma^2 + n_{ij} \cdot n \sigma_T^2$	$\frac{CM_{Trata}}{CM_{Error}}$
Rebañ(T)	R-t	$\sum_{ij} \frac{X_{ij.}^2}{n_{ij.}} - \sum_i \frac{X_{i..}^2}{n_{i..}}$	$\frac{SC_{Re}}{R-t}$	$\sigma^2 + n \sigma_R^2$	$\frac{CM_{Reba}}{CM_{Error}}$
Error	N-R	$\sum_{ijk} X_{ijk}^2 - \sum_{ij} \frac{X_{ij.}^2}{n_{ij.}}$	$\frac{SC_{Er}}{N-R}$	$\sigma^2$	
total	N-1	$\sum_{ijk} X_{ijk}^2 - \frac{X_{...}^2}{N}$			

Ejemplo.-

Resolvamos, ahora, los tres enunciados planteados en el epígrafe *Relación entre los factores*, que a su vez son adaptaciones de los enunciados del epígrafe *Factores de efectos fijos y factores de efectos aleatorios* y los tres epígrafes siguientes. Los datos para los tres enunciados eran los mismo, estos son

T <sub>1</sub>		T <sub>2</sub>		T <sub>3</sub>	
R <sub>1</sub>	R <sub>2</sub>	R <sub>1</sub>	R <sub>2</sub>	R <sub>1</sub>	R <sub>2</sub>
5	1	6	2	4	1
4	2	7	1	3	0
9	3	13	3	7	1
12		16		8	

El cálculo de los resultados comunes para los dos posibles análisis (correspondientes a los tres enunciados) son

$$\sum T^2 = \sum_i \frac{X_{i..}^2}{n_{i..}} = \frac{1}{4} (12^2 + 16^2 + 8^2) = 116$$

$$\sum R^2 = \sum_{ij} \frac{X_{ij.}^2}{n_{ij.}} = \frac{1}{2} (9^2 + 3^2 + 13^2 + 3^2 + 7^2 + 1^2) = 159$$

$$TC = \frac{X_{..}^2}{N} = \frac{36^2}{12} = 108$$

$$\sum X^2 = \sum_{ijk} X_{ijk}^2 = 5^2 + 4^2 + \dots + 1^2 + 0^2 = 162$$

Pasemos a resolver los tres enunciados, comenzando por los más corrientes, como son el enunciado 2 y el enunciado 3.

Enunciado 2: En este ejemplo, los tres niveles del factor principal son tres *tratamientos* que pueden ser tres dietas o tres antibióticos o tres tóxicos o tres métodos de purificación, *etc.*, es decir, un factor con tres únicos niveles del cual se quiere saber si existen diferencias entre dichos niveles y en caso positivo, saber cuál o cuales es el mejor. Con respecto al segundo factor, dos *rebaños* o dos hospitales o dos plantas de hospital o dos barrios o dos comarcas o dos ciudades *etc.*, no ocurre lo mismo sino que este segundo factor lo ha introducido el experimentador obligado por criterios ambientales o de clasificación. El experimentador no tiene interés en saber si existen diferencias entre los niveles de este segundo factor, se introduce en el modelo para explicar mejor la variabilidad observada. Como se ve, este ejemplo es semejante al Enunciado 2 del epígrafe *Factores de efectos fijos y factores de efectos aleatorios*, con la diferencia de que, aquí, un tratamiento se aplica en unos rebaños y otro tratamiento se aplicará a otros rebaños, aunque los rebaños serán semejantes; por tanto el modelo que lo explica es mixto, es decir, los tratamientos son fijos y los rebaños son aleatorios.

Enunciado 3: En este ejemplo, los tres niveles del factor principal o tres tratamientos pueden ser tres machos o tres familias, semejante al Enunciado 3 del epígrafe *Factores de efectos fijos y factores de efectos aleatorios*, por lo que serían, también, aleatorios, es decir, una muestra de la población de machos. Y el segundo factor puede seguir siendo dos *rebaños* o dos hospitales o dos plantas de hospital o dos barrios o dos comarcas o dos ciudades *etc.* Pero aquí, cada uno de los tres macho (los tres tratamientos) tiene descendencia en dos rebaños, que no son los mismos para los tres machos, con arreglo a la tabla de datos de este ejemplo.

En estos dos casos los análisis serían el mismo, este es.

Modelo Mixto y Aleatorio

<i>FV</i>	<i>gl</i>	<i>SC</i>	<i>CM</i>	<i>ECM</i>	<i>F<sub>o</sub></i>
<i>Trata</i>	3-1	116-108	$\frac{8}{2} = 4.00$	$\sigma^2 + 2 \sigma_R^2 + 4 \sigma_T^2$	$\frac{4.00}{14.33} = 0.279ns$
<i>Rebañ(T)</i>	6-3	159-116	$\frac{43}{3} = 14.33$	$\sigma^2 + 2 \sigma_R^2$	$\frac{14.33}{0.50} = 28.66***$
<i>Error</i>	12-6	162-159	$\frac{3}{6} = 0.5$	$\sigma^2$	
<i>total</i>	12-1	162-108			

Las *F* de contrastes son, para tratamiento,  $F_{(2,3; 0.05)}=9.55$  y, para rebaño dentro de tratamientos,  $F_{(3,6; 0.05)}=4.76$

Por lo que se concluye que los efectos de los tratamientos no son significativos. Esto significa, en el caso de tratamientos fijos, que no hay un tratamientos mejor que otro. Y en el caso de tratamientos aleatorios, significa que el efecto macho o familia, no influye significativamente en la variabilidad de la variable estudiada.

El factor jerárquico si influye en la variabilidad de la variable estudiada.

Enunciado 1: Aunque poco corriente en los modelos jerárquicos, puede ocurrir que el modelo hubiera sido fijo, es decir, tanto el factor  $T_i$  como el factor  $R_{i(j)}$  sean de efectos fijos, semejante al Enunciado 1 del epígrafe *Factores de efectos fijos y factores de efectos aleatorios*. El enunciado puede ser básicamente igual al anterior, esto es, los tres tratamientos puede ser tres dietas o tres antibióticos o tres ciudades o tres comarcas, y los dos niveles del factor jerárquico pueden ser dos rebaños (dentro de cada comarca, por tanto un total de seis rebaños) o dos hospitales o dos planta de hospital o dos centros de distribución (dentro de cada ciudad) siendo los tres tratamientos o las tres ciudades o las tres comarcas las únicas que nos interesan y los dos rebaños o dos hospitales los únicos que nos interesan.

En este caso el análisis sería este.

Modelo Fijo

<i>FV</i>	<i>gl</i>	<i>SC</i>	<i>CM</i>	<i>ECM</i>	<i>F<sub>o</sub></i>
<i>Trata</i>	3-1	116-108	$\frac{8}{2} = 4.00$	$\sigma^2 + 4\sigma_T^2$	$\frac{4.00}{0.5} = 8^*$
<i>Rebañ(T)</i>	6-3	159-116	$\frac{43}{3} = 14.33$	$\sigma^2 + 2\sigma_R^2$	$\frac{14.33}{0.50} = 28.66^{***}$
<i>Error</i>	12-6	162-159	$\frac{3}{6} = 0.5$	$\sigma^2$	
<i>total</i>	12-1	162-108			

Las *F* de contrastes son, para tratamiento,  $F_{(2,6; 0.05)}=5.14$  y, para rebaño dentro de tratamientos,  $F_{(3,6; 0.05)}=4.76$

Por lo que se concluye que los efectos de los tratamientos son diferentes, esto es hay uno o unos diferente de otro u otros tratamientos. Y lo mismo para el factor jerárquico; dentro de tratamientos.

**Archivo del programa SAS (C8-1.SAS).-**

```
options ls=75 ps=60;
title 'Análisis de la varianza, factores jerárquicos';
data jerarqui;
infile 'c8-1.dat';
input trata $ rebano $ prod @@;
title 'Modelo Fijo';
proc ANOVA;
  class trata rebano ;
  model prod = trata rebano(trata) ;
run;
title 'Modelo Mixto o Aleatorio';
  test H = trata E = rebano(trata);
run;
title 'Otra manera para el Modelo Mixto y el Aleatorio';
proc GLM;
  class trata rebano ;
  model prod = trata rebano(trata) ;
  random trata rebano(trata) / test ;
run;
proc sort; by trata rebano;run;
title 'Otra manera para el Modelo Mixto y el Aleatorio';
proc nested;
class trata rebano;
var prod;
run;
```

**Archivo de datos (C8-1.DAT).-**

T1 R1 5	T2 R1 6	T3 R1 4
T1 R1 4	T2 R1 7	T3 R1 3
T1 R2 1	T2 R2 2	T3 R2 1
T1 R2 2	T2 R2 1	T3 R2 0

**Archivo de resultados (C8-1.LST).-**

Modelo Fijo						
Dependent Variable: PROD						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	5	51.0000000	10.2000000	20.40	0.0011	
Error	6	3.0000000	0.5000000			
Corrected Total	11	54.0000000				
	R-Square	C.V.	Root MSE	PROD Mean		
	0.9444444	23.57023	0.70711	3.00000		
Source	DF	Anova SS	Mean Square	F Value	Pr > F	
TRATA	2	8.0000000	4.0000000	8.00	0.0203	
REBANO (TRATA)	3	43.0000000	14.3333333	28.67	0.0006	
Modelo Mixto o Aleatorio						
Dependent Variable: PROD						
Tests of Hypotheses using the Anova MS for REBANO (TRATA) as an error term						
Source	DF	Anova SS	Mean Square	F Value	Pr > F	
TRATA	2	8.0000000	4.0000000	0.28	0.7742	
Otra manera para el Modelo Mixto y el Aleatorio						
Dependent Variable: PROD						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	5	51.0000000	10.2000000	20.40	0.0011	
Error	6	3.0000000	0.5000000			
Corrected Total	11	54.0000000				
	R-Square	C.V.	Root MSE	PROD Mean		
	0.9444444	23.57023	0.70711	3.00000		
Source	DF	Type I SS	Mean Square	F Value	Pr > F	
TRATA	2	8.0000000	4.0000000	8.00	0.0203	
REBANO (TRATA)	3	43.0000000	14.3333333	28.67	0.0006	
Source	DF	Type III SS	Mean Square	F Value	Pr > F	
TRATA	2	8.0000000	4.0000000	8.00	0.0203	
REBANO (TRATA)	3	43.0000000	14.3333333	28.67	0.0006	
Source	Type III Expected Mean Square					
TRATA	Var(Error) + 2 Var(REBANO (TRATA)) + 4 Var (TRATA)					
REBANO (TRATA)	Var(Error) + 2 Var (REBANO (TRATA))					

Tests of Hypotheses for Random Model Analysis of Variance

Dependent Variable: PROD

Source: TRATA

Error: MS (REBANO (TRATA))

DF	Type III MS	Denominator DF	Denominator MS	F Value	Pr > F
2	4	3	14.33333333	0.2791	0.7742

Source: REBANO (TRATA)

Error: MS (Error)

DF	Type III MS	Denominator DF	Denominator MS	F Value	Pr > F
3	14.33333333	6	0.5	28.6667	0.0006

Otra manera para el Modelo Mixto y el Aleatorio  
Coefficients of Expected Mean Squares

Source	TRATA	REBANO	ERROR
TRATA	4	2	1
REBANO	0	2	1
ERROR	0	0	1

Nested Random Effects Analysis of Variance for Variable PROD

Variance Source	Degrees of Freedom	Sum of Squares	F Value	Pr > F	Error Term
TOTAL	11	54.000000			
TRATA	2	8.000000	0.279	0.7742	REBANO
REBANO	3	43.000000	28.667	0.0006	ERROR
ERROR	6	3.000000			

Variance Source	Mean Square	Variance Component	Percent of Total
TOTAL	4.909091	7.416667	100.0000
TRATA	4.000000	-2.583333	0.0000
REBANO	14.333333	6.916667	93.2584
ERROR	0.500000	0.500000	6.7416

Mean 3.00000000  
Standard error of mean 0.57735027

**Generalización del método para el análisis de modelos jerárquicos con más de dos factores.-**

Es importante conocer la mecánica del cálculo y análisis que debe seguirse para diversos tipos de modelos, pues aunque hasta ahora se han mostrado sus expresiones para el caso de jerárquicos con dos factores, interesa saber la generalización para cualquier número de factores.

En el caso de que se tenga un modelo jerárquico con  $f$  factores, el modelo sería

$$X_{ijk\dots f} = \mu + A_i + B_{j(i)} + C_{k(ij)} + \dots + F_{f(ijk\dots f-1)} + e_{ijk\dots f}$$

El esquema del experimento sería de este tipo

A <sub>1</sub>						A <sub>2</sub>					
B <sub>1</sub>		B <sub>2</sub>		B <sub>3</sub>		B <sub>1</sub>		B <sub>2</sub>		B <sub>3</sub>	
C <sub>1</sub>	C <sub>2</sub>	C <sub>1</sub>	C <sub>2</sub>	C <sub>1</sub>	C <sub>2</sub>	C <sub>1</sub>	C <sub>2</sub>	C <sub>1</sub>	C <sub>2</sub>	C <sub>1</sub>	C <sub>2</sub>

Nótese que  $B_1$  en  $A_1$  es diferentes que  $B_1$  en  $A_2$  aunque es semejante (ver el epígrafe *Diseños jerárquicos o con subgrupos*).

Este esquema es para un jerárquico con tres factores. Si hubiera más factores, habría que seguir descendiendo en el esquema (ver el esquema del jerárquico expuesto en el epígrafe *Es suficiente considerar dos factores*).

Nótese, así mismo, que los niveles del factor principal (en el caso concreto de la tabla anterior) es de dos, los niveles del segundo factor son seis o tres dentro de cada nivel del factor principal, y los niveles del tercer factor son doce o dos dentro de cada nivel del segundo factor.

El modelo lineal que explicaría esta experiencia sería

$$X_{ijkl} = \mu + A_i + B_{j(i)} + C_{k(ij)} + e_{ijkl}$$

El sistema para calcular los  $gl$  y  $SC$ , al igual que en los factoriales, tiene una lógica común; tan sólo hay que tener en cuenta que cuando se quiera hallar los grados de libertad, los números serán los niveles de los factores y para las sumas de cuadrados lo serán de los sumatorios correspondientes.

Los grados de libertad del factor principal (el primero), al igual que en el modelo de una vía, se obtienen restándole la unidad al número de niveles del factor considerado:  $a-1$ .

Para la suma de cuadrados del factor principal se resta del sumatorio del factor el término de corrección:  $\Sigma A^2 - TC$ .

Para los demás factores subordinados, tanto para el cálculo de los  $gl$  como para el cálculo de las  $SC$ , hay que restarle a dicho factor lo debido al factor inmediato superior, ya que parte de lo que se observa en aquél se debe a éste.

Resumamos lo dicho

<i>FV</i>	<i>gl</i>	<i>SC</i>
<i>A</i>	<i>a-1</i>	$\sum_i \frac{X_{i...}^2}{n_{i...}} - \frac{X_{....}^2}{N}$
<i>B(A)</i>	<i>b-a</i>	$\sum_{ij} \frac{X_{ij..}^2}{n_{ij..}} - \sum_i \frac{X_{i...}^2}{n_{i...}}$
<i>C(B)</i>	<i>c-b</i>	$\sum_{ijk} \frac{X_{ijk.}^2}{n_{ijk.}} - \sum_{ij} \frac{X_{ij..}^2}{n_{ij..}}$
<i>Error</i>	<i>N-c</i>	$\sum_{ijkl} X_{ijkl}^2 - \sum_{ijk} \frac{X_{ijk.}^2}{n_{ijk.}}$
<i>total</i>	<i>N-1</i>	$\sum_{ijkl} X_{ijkl}^2 - \frac{X_{....}^2}{N}$

*b* y *c* no es el número de niveles dentro del factor superior, sino que es el número total de niveles.

Una vez hallados los grados de libertad y las sumas de cuadrados, los cuadrados medios son las divisiones respectivas de las sumas de cuadrados por los grados de libertad, por lo que se esta en condiciones de probar las diferentes fuentes de variación. Para ello se tiene que saber cuales son los parámetros que estiman los cuadrados medios y, como se sabe, estos cambian con arreglo al tipo de factores que se tengan (fijos o aleatorios), por lo que se necesita un método que nos provea de estas estimas para saber, entre otras finalidades, cuales son las pruebas de significación para las diferentes fuentes de variación. Este método se verá más adelante, en el epígrafe *Cálculo de las estimas de los CM por el método rápido*.

Si bien es cierto que se puede afirmar que siempre que se tenga un jerárquico puro todos los factores serán aleatorios, exceptuando el principal que puede ser fijo o aleatorio. En cuyo caso se tendrá siempre un modelo aleatorio o mixto (siendo siempre el factor de efectos fijos el factor principal), por lo que la estimas de los cuadrados medios serán siempre las mismas, esto es, cada cuadrado medio estima, además de su varianza y la del error, la varianza del factor inmediatamente inferior, por lo que la *F* para la prueba de hipótesis será el cuadrado medio del factor en cuestión dividido por el cuadrado medio del factor inmediatamente inferior, hasta llegar al último que se divide por el error.

**Ejemplo.-**

Un ejemplo puede ser un estudio sobre la concentración de pigmento en el cabello humano (medido con un espectrofotómetro) realizado en 2 parejas



monozigóticas y 2 parejas dizigóticas. De cada persona se tomaron dos muestras de cabello, que se hidrolizaron por separado, y de cada muestra se realizó tres lecturas o medidas. Las lecturas o medidas con el espectrofotómetro fue completamente aleatoria. Los datos son

		1	2	3	4							
T P I M  i a n u p r d e o e i s j v t a i r d a u o	1	1	123	128	132	126	509	1017	3960			
		2	125	124	130	129				508		
	1	1	130	132	129	133	524	1046		2063		
		2	132	129	130	131					522	
	M	1	1	118	116	121	115	470		941	3960	
			2	117	116	118	120					471
		2	1	122	118	120	121	481		956		1897
			2	121	119	117	118					
	D	1	1	132	130	131	129	522		1031	3940	
			2	126	132	128	123					509
		1	1	131	130	129	132	522		1028		2059
			2	125	128	126	127					
2		1	1	122	111	120	115	468	943	1881		
			2	118	117	119	121					475
		2	1	115	121	116	118	470	938			1881
			2	119	120	114	115					

Este es un jerárquico puro de cuatro factores cuyo modelo lineal es

$$X_{ijklm} = \mu + T_i + P_{j(i)} + I_{k(ij)} + M_{l(ijk)} + e_{ijklm}$$

Los cálculos necesarios para el análisis son

$$\sum T^2 = \sum_i \frac{X_{i\dots}^2}{n_{i\dots}} = \frac{1}{32} (3960^2 + 3940^2) = 975162.5$$

$$\sum P^2 = \sum_{ij} \frac{X_{ij\dots}^2}{n_{ij\dots}} = \frac{1}{16} (2063^2 + 1997^2 + 2059^2 + 1881^2) = 977013.75$$

$$\sum I^2 = \sum_{ijk} \frac{X_{ijk\dots}^2}{n_{ijk\dots}} = \frac{1}{8} (1017^2 + 1046^2 + \dots + 943^2 + 938^2) = 977082.5$$

$$\sum M^2 = \sum_{ijkl} \frac{X_{ijkl}^2}{n_{ijkl}} = \frac{1}{4} (509^2 + 508^2 + \dots + 470^2 + 468^2) = 977147.5$$

$$TC = \frac{X_{\dots}^2}{N} = \frac{7900^2}{64} = 975156.25$$

$$\sum X^2 = \sum_{ijkln} X_{ijkln}^2 = 123^2 + 128^2 + \dots + 114^2 + 114^2 = 977466$$

Y el análisis es

<i>FV</i>	<i>gl</i>	<i>SC</i>	<i>CM</i>	<i>F<sub>o</sub></i>
<i>Tipo</i>	2-1	975162.5- 975156.25	$\frac{6.25}{1} = 6.25$	$\frac{6.25}{952.625} = 0.07ns$
<i>Pare(T)</i>	4-2	977013.75- 975162.5	$\frac{1851.25}{2} = 925.62$	$\frac{952.625}{17.1875} = 55.425^{**}$
<i>Indi(P)</i>	8-4	977082.5- 977013.75	$\frac{68.75}{4} = 17.1875$	$\frac{17.1875}{8.125} = 2.115ns$
<i>Mues(I)</i>	16-8	977147.5- 977082.5	$\frac{65}{8} = 8.125$	$\frac{8.125}{6.635} = 1.224ns$
<i>Error</i>	64-16	977466- 977147.5	$\frac{318.5}{48} = 6.6350$	
<i>total</i>	64-1	977466- 975156.25		

Las *F* de contrastes son, para **Tipo**,  $F_{(1,2; 0.05)}=18.5$ , para **Pareja** dentro de tipo,  $F_{(2,4; 0.05)}=6.94$ , para **Individuo** dentro de pareja,  $F_{(4,8; 0.05)}=3.84$  y para **Muestra** dentro de individuo,  $F_{(8,48; 0.05)}=2.14$

Por lo que se concluye que solo existen diferencias significativas entre parejas, es decir, que la concentración de pigmento en el cabello es diferente entre las diferentes parejas dentro del tipo de parejas. No existe diferencia entre el tipo de parejas (mellizos o gemelos) ni entre las personas dentro de la pareja, ni entre las muestras dentro de la persona.

### Archivo del programa SAS (C8-2.SAS) .-

```
title 'Jerárquico de cuatro factores';
options ls=75 ps=65;
data jerar;
infile 'c8-2.dat';
do tipo=1 to 2;
  do pareja= 1 to 2;
    do indivi=1 to 2;
      do muestra=1 to 2;
        do repe=1 to 4;
          input pigmento @@;output;
        end;
      end;
    end;
  end;
end;
proc sort;
by tipo pareja indivi muestra;
run;
proc nested;
  class tipo pareja indivi muestra;
  var pigmento;
run;
```

### Archivo de datos (C8-2.DAT) .-

```
123 128 132 126
125 124 130 129
130 132 129 133
132 129 130 131
118 116 121 115
117 116 118 120
122 118 120 121
121 119 117 118
132 130 131 129
126 132 128 123
131 130 129 132
125 128 126 127
122 111 120 115
118 117 119 121
115 121 116 118
119 120 114 115
```

### Archivo de resultados (C8-2.LST) .-

Jerárquico de cuatro factores					
Coefficients of Expected Mean Squares					
Source	TIPO	PAREJA	INDIVI	MUESTRA	ERROR
TIPO	32	16	8	4	1
PAREJA	0	16	8	4	1
INDIVI	0	0	8	4	1
MUESTRA	0	0	0	4	1
ERROR	0	0	0	0	1

Nested Random Effects Analysis of Variance for Variable PIGMENTO					
Variance Source	Degrees of Freedom	Sum of Squares	F Value	Pr > F	Error Term
TOTAL	63	2309.750000			
TIPO	1	6.250000	0.00675	0.9420	PAREJA
PAREJA	2	1851.250000	53.855	0.0013	INDIVI
INDIVI	4	68.750000	2.115	0.1705	MUESTRA
MUESTRA	8	65.000000	1.224	0.3055	ERROR
ERROR	48	318.500000			
Variance Source	Mean Square	Variance Component	Percent of Total		
TOTAL	36.662698	64.917969	100.0000		
TIPO	6.250000	-28.730469	0.0000		
PAREJA	925.625000	56.777344	87.4601		
INDIVI	17.187500	1.132813	1.7450		
MUESTRA	8.125000	0.372396	0.5736		
ERROR	6.635417	6.635417	10.2212		
Mean			123.43750000		
Standard error of mean			0.31250000		

### Bibliografía:

- Cochran, WG y Cox, GM.* 1965. DISEÑOS EXPERIMENTALES. Ed. TRILLAS. México.
- Li, Ching Chun.* 1969. INTRODUCCIÓN A LA ESTADÍSTICA EXPERIMENTAL. Ed. Omega. Barcelona.
- Lite, TM, y Jackson Hills, F.* 1987. MÉTODOS ESTADÍSTICOS PARA LA INVESTIGACIÓN EN LA AGRICULTURA. Ed. TRILLAS. México.
- Littel, R.C., Freund, R.J. and Spector, P.C.* 1991. SAS® FOR LINEAR MODELS. SAS Institute Inc., Cary, NC, USA.
- Ostle, B.* 1965. ESTADÍSTICA APLICADA. Ed. Limusa-Wiley. México.
- Ruiz-Maya, L.* 1983. MÉTODOS ESTADÍSTICOS DE INVESTIGACIÓN. Ed. INE. Madrid.
- Sokal, R.R. y Rohlf, F.J.* 1994. BIOMETRY. Ed. W.H.FREEMAN. San Francisco.
- Snedecor, G.W. y Cochran, W.G.* 1971. MÉTODOS ESTADÍSTICOS. Ed. C.E.C.S.A. México.
- Steel, R.* 1996. PRINCIPLES AND PROCEDURES OF STATISTICS. Ed. McGRAW-HILL Education. New York.
- SAS Institute Inc.* 1990. SAS/STAT® USER'S GUIDE. Volume 1 and 2. Cary, NC, USA.



## **CAPÍTULO 9**

# **Modelos Compuestos**



# CAPÍTULO 9

## Modelos Compuestos

### Modelos compuestos.-

Existe un tercer tipo de modelo que no se ha descrito hasta el momento, este es el que presenta conjuntamente efectos jerarquizados y efectos factoriales y al que se le denomina *Compuesto*.

Para continuar Veamos un ejemplo que además de presentar un caso de modelo compuesto puede ayudarnos a aclarar las diferencias entre un tipo y otro.

### Ejemplo.-

En un estudio se cuentan el número de insectos que aparecen en la época de floración y en la época de maduración del fruto, en dos años consecutivos, aplicando dos tratamientos. Los datos son los siguientes

<i>Año</i>	<i>Tratamiento</i>	<i>Época</i>	<i>Medida</i>
1	1	<i>floración</i>	$X_{111}$
		<i>maduración</i>	$X_{112}$
	2	<i>floración</i>	$X_{121}$
		<i>maduración</i>	$X_{122}$
2	1	<i>floración</i>	$X_{211}$
		<i>maduración</i>	$X_{212}$
	2	<i>floración</i>	$X_{221}$
		<i>maduración</i>	$X_{222}$

Como se ve hay tres factores que han podido influir en el número de insectos contados, estos son



## Año (A), Tratamiento (T) y Época (E)

Si no nos dan más datos, en este caso está claro que se trata de un triple factorial pues en los dos años y en las dos épocas se han realizado los dos tratamientos respectivos; por lo que el modelo sería

$$X_{ijkl} = \mu + A_i + T_j + E_k + AT_{ij} + AE_{ik} + TE_{jk} + ATE_{ijk} + e_{ijkl}$$

Pero podría haber ocurrido que las medidas tomadas fueran un conteo sobre plantas, por lo que habría que introducir, por tanto, este cuarto factor ( $P$ ), y según las plantas fueran anuales o perennes, el efecto de ese factor será uno u otro.

Si las plantas son anuales (por ejemplo, el algodón), el factor *planta* se encontrará jerarquizado respecto a *tratamiento* y a *año*, ya que las plantas del primer año son distintas de las del segundo año, y las plantas de un tratamiento son distintas de las del otro tratamiento, por lo que la expresión de su efecto sería:  $P_{l(ij)}$ . Evidentemente la planta es factorial con época pues cada planta estará representada en las dos épocas y para cada época estarán presentes todas las plantas. El modelo lineal, en este caso, sería

$$X_{ijklm} = \mu + A_i + T_j + E_k + P_{l(ij)} + AT_{ij} + AE_{ik} + TE_{jk} + EP_{kl(ij)} + ATE_{ijk} + e_{ijkl}$$

Si las plantas son perennes (por ejemplo, el olivo) tan sólo estarían jerarquizadas respecto a tratamiento (a no ser que, pudiendo ser, un tratamiento se aplique en una rama y el otro en otra rama) y no respecto a año, puesto que las plantas son las mismas a lo largo de los años y entonces se tendrá sólo  $P_{l(j)}$ , es decir, planta es factorial con época y con año pues cada planta estará representada en las dos épocas y años, y en cada época y año tenemos todas las plantas del experimento. El modelo sería

$$X_{ijklm} = \mu + A_i + T_j + E_k + P_{l(j)} + AT_{ij} + AE_{ik} + AP_{il(j)} + TE_{jk} + EP_{kl(j)} + ATE_{ijk} + AEP_{ikl(j)} + e_{ijklm}$$

### Más sobre las interacciones.-

Ahora que se han estudiado todos los tipos de modelos, insistamos en el concepto de interacción que ya se estudió, anteriormente, con los modelos factoriales o cruzados.

Recordemos que existe interacción entre dos factores cuando dentro de cada nivel de uno de ellos, los efectos de los diferentes niveles del otro factor no siguen la misma tendencia. Por ejemplo, si estamos analizando tres tratamientos en dos razas se puede encontrar que, para la  $R_1$ , el orden de los tratamientos de mejor a peor producción ha sido  $T_1 > T_3 > T_2$ , y en cambio en la  $R_2$  ha sido  $T_2 > T_3 > T_1$ , lo que quiere decir que entre tratamientos y razas existe una interacción y debe ser incluída en el Modelo, porque parte de la variabilidad total se deberá a dicha interacción.

Gráficamente se podría representar como ya se ha hecho anteriormente en el epígrafe *Interacción*.

Si un factor está jerarquizado respecto a otro principal, no se puede calcular la interacción entre los dos, puesto que, como se vio en el epígrafe *Relación entre los factores*, no se tiene todas las combinaciones de niveles entre ambos, es decir, no es un diseño cruzado o factorial. El que no se pueda calcular no indica nada sobre la existencia o no de dicha interacción, como se verá más adelante.

Las interacciones aparecen siempre que haya dos factores cruzados y no se puede hablar de interacción entre dos factores jerarquizados, por lo que hay que tenerlo en cuenta a la hora de establecer el modelo, que siempre es el primer paso para el estudio del diseño y análisis subsiguiente. Interesa ahora discutir lo que ocurrirá en un modelo compuesto de factores cruzados y jerarquizados.

a) Una serie de factores que están jerarquizados entre sí interaccionan con todos los que son independientes de esta subordinación.

b) Si, a su vez, este último tiene también factores subordinados, cada uno de ellos tendrá interacciones con los anteriores.

Por ejemplo

$$X_{ijkl} = \mu + A_i + B_{j(i)} + C_{k(ij)} + D_l + AD_{il} + BD_{j(i)l} + CD_{k(ij)l} + e_{ijkl}$$

Otro ejemplo

$$X_{ijkl} = A_i + B_{j(i)} + C_{k(ij)} + D_l + E_{m(l)} + AD_{il} + BD_{j(i)l} + CD_{k(ij)l} + AE_{im(l)} + BE_{jm(i)l} + CE_{km(ij)l} + e_{ijkl}$$

Hagamos una pequeña discusión de la posibilidad de existencia y la posibilidad de cálculo de las interacciones.

Es evidente que se pueden dar casos en que no tenga sentido dicha interacción por imposibilidad biológica, física, etc. Por ejemplo, si se está probando unos tratamientos con dos razas distintas y dentro de cada raza se ha tomado, por ejemplo, 3 lotes de animales (bloques), esta claro que el efecto bloque está jerarquizado respecto al de raza por lo que no se podría, de ninguna manera, hablar de interacción *raza/lote* pues biológicamente es imposible que un lote de animales (bloque) de una raza lo supongamos después de otra raza distinta para poder contrastar esa hipotética interacción.

Puede haber casos en que, no existiendo dicha incompatibilidad conceptual, diseñemos una experiencia jerarquizada en la que no se podrá medir la interacción. Esto puede deberse a; 1) que sabemos que no existe dicha interacción, por el tipo de factores de que se trata o por experiencias anteriores; 2) aun reconociendo que puede existir la interacción, nos es indiferente y no la consideramos, por no afectar a nuestro problema o por ser excesivamente cara o difícil de diseñar una experiencia factorial. Es

evidente que en este último caso las conclusiones deberán ser cautelosas respecto a las hipótesis planteadas.

Un caso típico es el de apareamiento para la selección de gallinas, o de sementales de carne en vacuno, por ejemplo, en el que se cruza cada macho con varias hembras, quedando por tanto un modelo jerarquizado de padres, madres e hijas, no se puede introducir la interacción en el modelo lineal, sin embargo se puede utilizar la idea de una interacción entre los efectos de padre y madre puesto que la descendencia de un macho con una hembra determinada puede ser excelente y en cambio con otra hembra puede dar unas hijas mediocres, pudiendo ser ésta muy buena madre si fuera apareada con otro macho; y se puede utilizar esta idea de interacción (sin introducirla en el modelo lineal) debido a los conocidos fenómenos genéticos de *epistasia* y *dominancia*. Un diseño de tipo *dialélico* (factorial) en el que se cruzan todos los machos con todas las hembras, para tener una descendencia en la que se conociera su genealogía, exigiría realizar tandas sucesivas y espaciadas en el tiempo de forma que los machos fueran apareándose de una manera rotativa. Esto, aparte de muy complicado y costoso introduce un nuevo factor temporal o estacional que no permitiría, en principio, analizar bien dichas interacciones; pero lo más importante es que el diseño jerarquizado permite realizar las estimas de los parámetros buscados sin que haya que tener en cuenta para nada las hipotéticas interacciones.

### **Generalización del método para el análisis de Modelos Compuestos.-**

Es importante conocer la mecánica del cálculo y análisis que debe seguirse en cualquier tipo de modelo. Hasta ahora se han mostrado los métodos de cálculo para la generalidad de los modelos factoriales o jerárquicos. Interesa saber la generalización para el tercer, y último, tipo de modelos, esto es, para los *Modelos Compuestos*.

Recordemos que en los modelos compuestos los factores jerárquicos pueden serlo respecto a otro factor o respecto a una combinación de varios. Así mismo, como ya se ha dicho anteriormente, un efecto factorial se puede cruzar con otro factorial (o principal) o con un factor jerárquico.

El cálculo de los grados de libertad y de las sumas de cuadrados se realiza siguiendo las mismas reglas para la generalización que hemos dado tanto para los diseños factoriales como para los jerárquicos. Si bien vamos a repetir las para los dos casos que no hemos citado anteriormente, si bien, insistimos, las reglas son las mismas.

### **Modelo compuesto con un jerárquico dentro de una interacción.-**

Este puede ser

$$X_{ijkl} = \mu + A_i + B_j + AB_{ij} + C_{k(ij)} + e_{ijkl}$$

que respondería a un esquema experimental de este tipo

	A <sub>1</sub>		A <sub>2</sub>	
	C <sub>1</sub>	C <sub>2</sub>	C <sub>1</sub>	C <sub>2</sub>
B <sub>1</sub>	n <sub>11(1)1</sub>	n <sub>11(2)1</sub>	n <sub>21(3)1</sub>	n <sub>21(4)1</sub>
	n <sub>11(1)2</sub>	n <sub>11(2)2</sub>	n <sub>21(3)2</sub>	n <sub>21(4)2</sub>
B <sub>2</sub>	n <sub>12(5)1</sub>	n <sub>12(6)1</sub>	n <sub>22(7)1</sub>	n <sub>22(8)1</sub>
	n <sub>12(5)2</sub>	n <sub>12(6)2</sub>	n <sub>22(7)2</sub>	n <sub>22(8)2</sub>
B <sub>3</sub>	n <sub>13(9)1</sub>	n <sub>13(10)1</sub>	n <sub>23(11)1</sub>	n <sub>23(12)1</sub>
	n <sub>13(9)2</sub>	n <sub>13(10)2</sub>	n <sub>23(11)2</sub>	n <sub>23(12)2</sub>

Nótese que B<sub>1</sub> en A<sub>1</sub> es el mismo nivel que en A<sub>2</sub>, mientras que C<sub>1</sub> en A<sub>1</sub>B<sub>1</sub> es diferente que en A<sub>1</sub>B<sub>2</sub>.

Nótese, así mismo, que los niveles de A son dos, los de B son tres y los de C son doce, esto es, dos dentro de cada casilla de AB.

El sistema para calcular los *gl* y *SC* de los factores principales y de la interacción, se realiza igual que se vio con los factoriales puros y los jerárquicos puros. Pero la lógica para los grados de libertad y las sumas de cuadrados del jerárquico dentro de interacción es la misma, como se va a ver a continuación.

Tanto para el cálculo de los *gl* como para el cálculo de las *SC* del jerárquico dentro de interacción, hay que restarle lo debido al factor jerárquico, lo debido al factor inmediato superior, que en este caso es la interacción, ya que parte de lo que se observa en aquél se debe a ésta. Es decir,  $gl = c - ab$ , en lo relativo a los grados de libertad y  $\Sigma C^2 - \Sigma AB^2$ , para la suma de cuadrados.

El resumen del análisis de este modelo sería

FV	<i>gl</i>	SC
A	a-1	$\sum_i \frac{X_{i...}^2}{n_{i...}} - \frac{X_{....}^2}{N}$
B	b-1	$\sum_j \frac{X_{.j..}^2}{n_{.j..}} - \frac{X_{....}^2}{N}$
AB	(a-1)(b-1)	$\sum_{ij} \frac{X_{ij..}^2}{n_{ij..}} - \sum_i \frac{X_{i...}^2}{n_{i...}} - \sum_j \frac{X_{.j..}^2}{n_{.j..}} + \frac{X_{....}^2}{N}$
C	c-ab	$\sum_{ijk} \frac{X_{ijk.}^2}{n_{ijk.}} - \sum_{ij} \frac{X_{ij..}^2}{n_{ij..}}$

Error	N-c	$\sum_{ijkl} X_{ijkl}^2 - \sum_{ijk} \frac{X_{ijk}^2}{n_{ijk}}$
Total	N-1	$\sum_{ijkl} X_{ijkl}^2 - \frac{X^2}{N}$

c no es el número de niveles dentro del factor superior, la interacción, sino que es el número total de niveles.

Una vez hallados los grados de libertad y las sumas de cuadrados, los cuadrados medios son las divisiones respectivas de las sumas de cuadrados por los grados de libertad, por lo que se esta en condiciones de probar las diferentes fuentes de variación. Para ello se tiene que saber cuales son los parámetros que estiman los cuadrados medios y, como se sabe, estos cambian con arreglo al tipo de factores que se tengan (fijos o aleatorios), por lo que se necesita un método que nos provea de estas estimas para saber, entre otras finalidades, cuales son las pruebas de significación para las diferentes fuentes de variación. Este método se verá en el siguiente epígrafe, *Cálculo de las estimas de los CM por el método rápido*.

**Modelo compuesto con una interacción de un factor principal con un factor jerárquico.-**

Puede ser

$$X_{ijkl} = \mu + A_i + B_{j(i)} + C_k + AC_{ik} + BC_{j(i)k} + e_{ijkl}$$

que respondería a un esquema experimental de este tipo

	A <sub>1</sub>			A <sub>2</sub>		
	B <sub>1</sub>	B <sub>2</sub>	B <sub>3</sub>	B <sub>1</sub>	B <sub>2</sub>	B <sub>3</sub>
C <sub>1</sub>	$n_{1(1)11}$	$n_{1(2)11}$	$n_{1(3)11}$	$n_{2(4)11}$	$n_{2(5)11}$	$n_{2(6)11}$
	$n_{1(1)12}$	$n_{1(2)12}$	$n_{1(3)12}$	$n_{2(4)12}$	$n_{2(5)12}$	$n_{2(6)12}$
C <sub>2</sub>	$n_{1(1)21}$	$n_{1(2)21}$	$n_{1(3)21}$	$n_{2(4)21}$	$n_{2(5)21}$	$n_{2(6)21}$
	$n_{1(1)22}$	$n_{1(2)22}$	$n_{1(3)22}$	$n_{2(4)22}$	$n_{2(5)22}$	$n_{2(6)22}$

Nótese que B<sub>1</sub> en A<sub>1</sub> es diferente que B<sub>1</sub> en A<sub>2</sub>, mientras que C<sub>1</sub> en A<sub>1</sub> es el mismo nivel que en A<sub>2</sub>.

Nótese, así mismo, que los niveles de A son dos, los de B son seis o tres dentro de cada nivel de A y los de C son dos.

Tanto para el cálculo de los *gl* como para el cálculo de las *SC*, de la interacción del *factor principal x jerárquico*, se multiplican los grados de libertad y las sumas de cuadrados, respectivamente, de ambos factores. Es decir,  $(b-a)(c-1) = bc - ac - b + a$  en lo

relativo a los grados de libertad y  $(\Sigma B^2 - \Sigma A^2)(\Sigma C^2 - 1) = \Sigma BC^2 - \Sigma AC^2 - \Sigma B^2 + \Sigma A^2$ , para la suma de cuadrados.

El resumen del análisis de este modelo sería

<i>FV</i>	<i>gl</i>	<i>SC</i>
<i>A</i>	<i>a-1</i>	$\sum_i \frac{X_{i...}^2}{n_{i...}} - \frac{X_{...}^2}{N}$
<i>B(A)</i>	<i>b-a</i>	$\sum_{ij} \frac{X_{ij..}^2}{n_{ij..}} - \sum_i \frac{X_{i...}^2}{n_{i...}}$
<i>C</i>	<i>a-1</i>	$\sum_k \frac{X_{..k.}^2}{n_{..k.}} - \frac{X_{...}^2}{N}$
<i>AC</i>	<i>ac-a-c+1</i>	$\sum_{ik} \frac{X_{i.k.}^2}{n_{i.k.}} - \sum_i \frac{X_{i...}^2}{n_{i...}} - \sum_k \frac{X_{..k.}^2}{n_{..k.}} + \frac{X_{...}^2}{N}$
<i>B(A)C</i>	<i>bc-ac-b+a</i>	$\sum_{jk} \frac{X_{.jk.}^2}{n_{.jk.}} - \sum_{ik} \frac{X_{i.k.}^2}{n_{i.k.}} - \sum_j \frac{X_{.j..}^2}{n_{.j..}} + \sum_i \frac{X_{i...}^2}{n_{i...}}$
<i>Error</i>	<i>N-bc</i>	$\sum_{ijkl} X_{ijkl}^2 - \sum_{jk} \frac{X_{.jk.}^2}{n_{.jk.}}$
<i>total</i>	<i>N-1</i>	$\sum_{ijkl} X_{ijkl}^2 - \frac{X_{...}^2}{N}$

*b* no es el número de niveles dentro del factor superior, el *A*, sino que es el número total de niveles.

Una vez hallados los grados de libertad y las sumas de cuadrados, los cuadrados medios son las divisiones respectivas de las sumas de cuadrados por los grados de libertad, por lo que se esta en condiciones de probar las diferentes fuentes de variación. Para ello se tiene que saber cuales son los parámetros que estiman los cuadrados medios y, como se sabe, estos cambian con arreglo al tipo de factores que se tengan (fijos o aleatorios), por lo que se necesita un método que nos provea de estas estimas para saber, entre otras finalidades, cuales son las pruebas de significación para las diferentes fuentes de variación. Este método se verá en el siguiente epígrafe, *Cálculo de las estimas de los CM por el método rápido.*

**Cálculo de las estimas de los CM por el método rápido.-**

Cómo el cálculo analítico de las estimas de los *CM*, explicado para modelos más sencillos, resulta largo y tedioso, vamos a exponer aquí un tipo de algoritmo, debido a *Franklin y Bennet*, que simplifica su obtención en un caso general.

Sea el Modelo

$$X_{ijkl} = \mu + A_i + B_{j(l)} + C_k + AC_{ik} + BC_{kj(l)} + e_{ijkl}$$

al que le corresponde el siguiente esquema experimental

	$A_1$		$A_2$		$A_3$	
	$B_1$	$B_2$	$B_3$	$B_4$	$B_5$	$B_6$
$C_1$						
$C_2$						
$C_3$						
$C_4$						

Considérese que

$A$  es de efecto fijo  
 $B$  " " " aleatorio  
 $C$  " " " fijo

Los niveles son

- i) 1, 2, 3
- j) 1, 2 dentro de cada  $i$
- k) 1, 2, 3, 4
- l) 1, ...,  $n$  dentro de cada  $ijk$

Formamos un cuadro de doble entrada de la forma siguiente: Se hacen tantas columnas como factores principales haya, contando el término error; cada columna va encabezada con el número de niveles del factor, el subíndice que le corresponde y la indicación de si es fijo o aleatorio; en los jerárquicos se indica el número de niveles dentro de cada orden superior.

En las filas se consideran todos los sumandos que intervienen en el modelo empezando por el último y sin tener en cuenta la media.

	3	2	4	$n$	
	$F$	$A$	$A$	$A$	$ECM$
	$i$	$j$	$k$	$l$	
$e_{ijkl}$	1	1	1	1	$\sigma^2$
$BC_{ij(l)}$	1	1	0	$n$	$\sigma^2 + n\sigma_{BC}^2$
$AC_{ik}$	0	2	0	$n$	$\sigma^2 + n\sigma_{BC}^2 + 2n\sigma_{AC}^2$
$C_k$	3	2	0	$n$	$\sigma^2 + n\sigma_{BC}^2 + 6n\sigma_C^2$
$B_{l(l)}$	1	1	4	$n$	$\sigma^2 + 4n\sigma_B^2$
$A_i$	0	2	4	$n$	$\sigma^2 + 4n\sigma_B^2 + 8n\sigma_A^2$

Los cuadros de la parte central, se rellenan yendo por columnas y la cifra que se pone en la intersección con cada fila viene determinada de la siguiente forma

- si la fila tiene el subíndice, de la columna de que se trate, entre paréntesis, se pone un 1.
- si tiene dicho subíndice fuera del paréntesis, se pone un 1 si hay una  $A$  en la cabeza de la columna y un 0 si hay una  $F$ .
- si en la fila no está el subíndice se pone el número de niveles de la columna.

Las  $ECM$  se van escribiendo por filas. En cada una hay un polinomio que consta de varios sumandos con diferentes varianzas afectadas por unos coeficientes

- el polinomio de cada fila contendrá la varianza del efecto de dicha fila más las varianzas de los efectos de las filas anteriores (no posteriores) que contengan en sus subíndices los subíndices de la fila que se está considerando.
- se tapan (o lo que es lo mismo, no se tienen en cuenta) las columnas en donde figuran subíndices del efecto de la fila considerada.
- a cada varianza que haya que incluir se le pone como coeficiente el producto de los números que queden en la fila de su correspondiente efecto. Naturalmente, si ese producto fuera cero desaparece del polinomio dicha varianza, aunque ésta fuera de las que en teoría tendría que incluirse.

Todo este proceso es válido para todas las filas, pero en la que contiene el término error se puede colocar directamente  $\sigma^2$ , y este irá en todos los polinomios.



**Ejemplo.-**

Un etólogo piensa que el grado de inteligencia de las ratas depende de la influencia de la madre y/o hermanos de camada durante el desarrollo. Para probar esta hipótesis toma una muestra aleatoria, del animalario, de cuatro ratas de las que obtiene dos camadas. De cada camada toma aleatoriamente tres grupos de tres ratas recién nacidas; un grupo es criado con la madre, otro grupo es criado sin la madre pero manteniendo juntos los hermanos, y los individuos del tercer grupo son criados completamente separados.

La variable que mide es la longitud del recorrido que realizan las ratas en un laberinto diseñado al efecto.

		Sistema de crianza								
		Con madre Con hermanos			Sin madre Con hermanos			Sin madre Sin hermanos		
Madre 1	Camada 1	1.7	1.9	1.7	1.8	1.9	2.0	2.5	2.6	2.3
	Camada 2	1.2	1.8	1.4	1.9	1.8	1.6	2.5	2.7	2.7
Madre 2	Camada 1	0.9	1.0	1.1	1.5	1.5	1.6	2.1	2.4	2.2
	Camada 2	0.8	0.8	1.0	1.5	1.6	1.4	2.0	2.3	2.2
Madre 3	Camada 1	1.4	1.3	1.3	1.7	1.7	1.8	2.3	2.1	2.2
	Camada 2	1.2	1.4	1.4	1.6	1.7	1.8	2.0	2.2	2.4
Madre 4	Camada 1	1.3	1.6	1.7	1.7	1.9	1.6	2.3	2.5	2.5
	Camada 2	1.4	1.7	1.8	1.7	1.7	1.5	2.6	2.4	2.6

Lo primero es determinar cuáles son los factores, el tipo de factores, la relación entre ellos y el número de niveles de cada uno

Factores	Símbolo	Tipo	Relación	Niveles
Sistema	(S)	Fijo	↕	$i=1,2,3$
Madre	(M)	Aleatorio	↕	$j=1,2,3,4$
Camada	(C)	Aleatorio	↕	$k=2 \times 4$
Individuo	(e)	Aleatorio	↕	$l=1,2,3$

siendo  $\begin{matrix} a \uparrow \\ b \downarrow \end{matrix}$  jerarquización de  $b$  a  $a$ , y  $\begin{matrix} a \uparrow \\ b \downarrow \end{matrix}$  factores cruzados

El modelo lineal será por tanto

$$X_{ijkl} = \mu + S_i + M_j + C_{k(j)} + SM_{ij} + SC_{ik(j)} + e_{ijkl}$$

Las estimas de los cuadrados medios son

	3 <i>F</i> <i>i</i>	4 <i>A</i> <i>j</i>	2 <i>A</i> <i>k</i>	3 <i>A</i> <i>l</i>	ECM
$e_{ijkl}$	1	1	1	1	$\sigma^2$
$SC_{ik(l)}$	0	1	1	3	$\sigma^2 + 3\sigma_{SC}^2$
$SM_{ij}$	0	1	2	3	$\sigma^2 + 3\sigma_{SC}^2 + 6\sigma_{SM}^2$
$C_{k(l)}$	3	1	1	3	$\sigma^2 + 9\sigma_C^2$
$M_j$	3	1	2	3	$\sigma^2 + 9\sigma_C^2 + 18\sigma_M^2$
$S_i$	0	4	2	3	$\sigma^2 + 3\sigma_{SC}^2 + 6\sigma_{SM}^2 + 24\sigma_S^2$

Las sumas de cuadrados y los grados de libertad del análisis de varianza serían

FV	gl	SC
Sistema	$s-1$	$\sum_i \frac{X_{i...}^2}{n_{i...}} - \frac{X^2}{N}$
Madre	$m-1$	$\sum_j \frac{X_{.j..}^2}{n_{.j..}} - \frac{X^2}{N}$
Cama(Madr)	$c-m$	$\sum_{ijk} \frac{X_{ijk.}^2}{n_{ijk.}} - \sum_j \frac{X_{.j..}^2}{n_{.j..}}$
Sistx Madr	$(s-1)(m-1)$	$\sum_{ij} \frac{X_{ij..}^2}{n_{ij..}} - \sum_i \frac{X_{i...}^2}{n_{i...}} - \sum_j \frac{X_{.j..}^2}{n_{.j..}} + \frac{X^2}{N}$
Sisx Ca(M)	$(s-1)(c-m)$	$\sum_{ik} \frac{X_{i.k.}^2}{n_{i.k.}} - \sum_{ij} \frac{X_{ij..}^2}{n_{ij..}} - \sum_{ijk} \frac{X_{ijk.}^2}{n_{ijk.}} + \sum_j \frac{X_{.j..}^2}{n_{.j..}}$
Error	$N-sc$	$\sum_{ijkl} X_{ijkl}^2 - \sum_{ik} \frac{X_{i.k.}^2}{n_{i.k.}}$
total	$N-1$	$\sum_{ijkl} X_{ijkl}^2 - \frac{X^2}{N}$

Y las pruebas  $F$  serían

<i>FV</i>	<i>ECM</i>	<i>F<sub>o</sub></i>
<i>Sistema</i>	$\sigma^2 + 3\sigma_{SC}^2 + 6\sigma_{SM}^2 + 24\sigma_S^2$	$\frac{CM_S}{CM_{SM}}$
<i>Madre</i>	$\sigma^2 + 9\sigma_C^2 + 18\sigma_M^2$	$\frac{CM_M}{CM_{C(M)}}$
<i>Cama(Madre)</i>	$\sigma^2 + 9\sigma_C^2$	$\frac{CM_{C(M)}}{CM_{Error}}$
<i>Sist xMadre</i>	$\sigma^2 + 3\sigma_{SC}^2 + 6\sigma_{SM}^2$	$\frac{CM_{SM}}{CM_{SC(M)}}$
<i>Sis x Ca(M)</i>	$\sigma^2 + 3\sigma_{SC}^2$	$\frac{CM_{SC(M)}}{CM_{Error}}$
<i>Error</i>	$\sigma^2$	
<i>Total</i>		

Los cálculos necesarios para el análisis son

$$\sum S^2 = \sum_i \frac{X_{i...}^2}{n_{i...}} = \frac{1}{24} (32.8^2 + 40.5^2 + 56.6^2) = 246.6521$$

$$\sum M^2 = \sum_j \frac{X_{.j..}^2}{n_{.j..}} = \frac{1}{18} (36.0^2 + 27.9^2 + 31.5^2 + 34.5^2) = 236.495$$

$$\sum C^2 = \sum_{ijk} \frac{X_{ijk.}^2}{n_{ijk.}} = \frac{1}{9} (18.4^2 + 17.6^2 + \dots + 17.1^2 + 17.4^2) = 236.5633$$

$$\sum SM^2 = \sum_{ij} \frac{X_{ij..}^2}{n_{ij..}} = \frac{1}{6} (9.7^2 + 11^2 + \dots + 10.1^2 + 14.9^2) = 249.365$$

$$\sum SC^2 = \sum_{ik} \frac{X_{i.k.}^2}{n_{i.k.}} = \frac{1}{3} (5.3^2 + 5.7^2 + \dots + 4.9^2 + 7.6^2) = 249.65$$

$$TC = \frac{X_{...}^2}{N} = \frac{129.9^2}{72} = 234.3613$$

$$\sum X^2 = \sum_{ijkl} X_{ijkl}^2 = 1.7^2 + 1.9^2 + \dots + 2.4^2 + 2.6^2 = 250.63$$

Y el análisis es

<i>FV</i>	<i>gl</i>	<i>SC</i>	<i>CM</i>	<i>F<sub>o</sub></i>
<i>Sistema</i>	3-1	246.6521-234.3613	$\frac{12.2908}{2} = 6.1454$	$\frac{6.1454}{0.0965} = 63.68^{***}$
<i>Madre</i>	4-1	236.495-234.3613	$\frac{2.1337}{3} = 0.7112$	$\frac{0.7112}{0.0171} = 41.59^{**}$
<i>Cam(Ma)</i>	8-4	236.5633-236.495	$\frac{0.0683}{4} = 0.0171$	$\frac{0.0171}{0.0204} = 0.84sn$
<i>Sis Mad</i>	2x3	249.365-246.6521- 236.495+234.3613	$\frac{0.5792}{6} = 0.0965$	$\frac{0.0965}{0.0271} = 3.56ns$
<i>Si Ca(M)</i>	2x4	249.65-249.365- 236.5633+236.495	$\frac{0.2167}{8} = 0.0271$	$\frac{0.0271}{0.0204} = 1.33ns$
<i>Error</i>	72-3x8	250.6301-249.65	$\frac{0.9801}{48} = 0.0204$	
<i>total</i>	72-1	250.6301-234.3613		

Las *F* de contrastes son, para *Sistema*,  $F_{(2,6; 0.05)}=5.14$ ; para *Madre*,  $F_{(3,4; 0.05)}=6.59$ ; para *Camada dentro de madre*,  $F_{(4,48; 0.05)}=2.57$ ; para la interacción *Sistema/Madre*,  $F_{(6,8; 0.05)}=3.58$ ; y para la interacción *Sistema/Camada dentro de Madre*,  $F_{(8,48; 0.05)}=2.14$ .

Por lo que se concluye que, en el grado de inteligencia de las ratas, influye tanto el sistema de crianza como la madre, no siendo significativa la influencia de la camada ni ninguna de las dos posibles interacciones.

### Archivo del programa SAS (C9-1.SAS) .-

```

title 'Análisis de la varianza Compuesto';
option ls=75 ps=65;
data biome;
infile 'C9-1.dat';
do madre = 1 to 4;
  do camada = 1 to 2;
    do sistema = 1 to 3;
      do indivi = 1 to 3;
        input inteli @@;output;
      end;
    end;
  end;
end;
proc glm;
  class sistema madre camada;
  model inteli = sistema madre camada(madre)
            sistema*madre sistema*camada(madre);
  test H=sistema          E=sistema*madre;
  test H=madre            E=camada(madre);
  test H=sistema*madre    E=sistema*camada(madre);
run;

```

### Archivo de datos (C9-1.DAT) .-

1.7	1.9	1.7	1.8	1.9	2.0	2.5	2.6	2.3
1.2	1.8	1.4	1.9	1.8	1.6	2.5	2.7	2.7
0.9	1.0	1.1	1.5	1.5	1.6	2.1	2.4	2.2
0.8	0.8	1.0	1.5	1.6	1.4	2.0	2.3	2.2
1.4	1.3	1.3	1.7	1.7	1.8	2.3	2.1	2.2
1.2	1.4	1.4	1.6	1.7	1.8	2.0	2.2	2.4
1.3	1.6	1.7	1.7	1.9	1.6	2.3	2.5	2.5
1.4	1.7	1.8	1.7	1.7	1.5	2.6	2.4	2.6

### Archivo de resultados (C9-1.LST) .-

Análisis de la varianza Compuesto						
Dependent Variable: INTELI						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	23	15.2887500	0.6647283	32.56	0.0001	
Error	48	0.9800000	0.0204167			
Corrected Total	71	16.2687500				
	R-Square	C.V.	Root MSE	INTELI Mean		
	0.939762	7.919828	0.14289	1.80417		
Source	DF	Type I SS	Mean Square	F Value	Pr > F	
SISTEMA	2	12.2908333	6.1454167	301.00	0.0001	
MADRE	3	2.1337500	0.7112500	34.84	0.0001	
CAMADA (MADRE)	4	0.0683333	0.0170833	0.84	0.5087	
SISTEMA*MADRE	6	0.5791667	0.0965278	4.73	0.0007	
SISTEMA*CAMADA (MADRE)	8	0.2166667	0.0270833	1.33	0.2535	

Source	DF	Type III SS	Mean Square	F Value	Pr > F
SISTEMA	2	12.2908333	6.1454167	301.00	0.0001
MADRE	3	2.1337500	0.7112500	34.84	0.0001
CAMADA (MADRE)	4	0.0683333	0.0170833	0.84	0.5087
SISTEMA*MADRE	6	0.5791667	0.0965278	4.73	0.0007
SISTEM*CAMADA (MADRE)	8	0.2166667	0.0270833	1.33	0.2535

Tests of Hypotheses using the Type III MS for SISTEMA\*MADRE as an error term

Source	DF	Type III SS	Mean Square	F Value	Pr > F
SISTEMA	2	12.2908333	6.1454167	63.66	0.0001

Tests of Hypotheses using the Type III MS for CAMADA (MADRE) as an error term

Source	DF	Type III SS	Mean Square	F Value	Pr > F
MADRE	3	2.13375000	0.71125000	41.63	0.0018

Tests of Hypotheses using the Type III MS for SISTEM\*CAMADA (MADRE) as an error term

Source	DF	Type III SS	Mean Square	F Value	Pr > F
SISTEMA*MADRE	6	0.57916667	0.09652778	3.56	0.0506

Otro ejemplo que daría el mismo modelo es el siguiente.

### Ejemplo.-

Se quiere saber si existen diferencias en las mediciones, realizadas por dos cardiólogos, del volumen máximo del ventrículo izquierdo durante el ciclo cardíaco, a partir de mediciones hechas en radiografías del tórax. Con esta finalidad se realizan dos radiografías de cada uno de diez pacientes. Cada radiografía es utilizada dos veces por cada uno de los dos médicos para obtener dos medidas, dado que los contornos radiográficos no son nítidos. Las medidas se realizan de modo que el médico no sepa a que radiografía corresponde, esto es, el médico no sabe a que paciente corresponde ni sabe si es la primera o la segunda vez que la mide. Por lo que cada médico analiza dos veces cada una de dos radiografías de cada uno de diez pacientes.

Si representáramos los datos, estos estarían en una tabla semejante al del experimento anterior, donde el factor *Sistema de Crianza* sería ahora el factor *Médico*, con dos niveles en lugar de tres; el factor *Madre* sería ahora el factor *Paciente*, con diez niveles en lugar de cuatro; y el factor *Camada dentro de Madre* sería ahora el factor *Radiografía dentro de Paciente* con dos niveles dentro de cada paciente, por lo tanto, con veinte niveles en total. Las repeticiones en las medidas constituirán la variabilidad individual o error.

Se tiene, por lo tanto

Factores	Símbolo	Tipo	Relación	Niveles
Médico	(M)	Fijo	↕	$i=2$
Paciente	(P)	Aleatorio	↑	$j=10$
Radiografía	(R)	Aleatorio	↕	$k=2 \times 10$
Repetición	(e)	Aleatorio	↑	$l=2$

El modelo lineal será por tanto

$$X_{ijkl} = \mu + M_i + P_j + R_{k(j)} + MP_{ij} + MR_{ik(j)} + e_{ijkl}$$

Las estimas de los cuadrados medios son

	2	10	2	2	
	F	A	A	A	
	i	j	k	l	ECM
$e_{ijkl}$	1	1	1	1	$\sigma^2$
$MR_{ik(j)}$	0	1	1	2	$\sigma^2 + 2\sigma_{MR}^2$
$MP_{ij}$	0	1	2	2	$\sigma^2 + 2\sigma_{MR}^2 + 4\sigma_{MP}^2$
$R_{k(j)}$	2	1	1	2	$\sigma^2 + 4\sigma_R^2$
$P_j$	2	1	2	2	$\sigma^2 + 4\sigma_R^2 + 8\sigma_P^2$
$M_i$	0	10	2	2	$\sigma^2 + 2\sigma_{MR}^2 + 4\sigma_{MP}^2 + 40\sigma_M^2$

Las sumas de cuadrados y los grados de libertad del análisis de varianza serían

<i>FV</i>	<i>gl</i>	<i>SC</i>
<i>Médico</i>	1	$\sum_i \frac{X_{i...}^2}{n_{i...}} - \frac{X_{....}^2}{N}$
<i>Paciente</i>	9	$\sum_j \frac{X_{.j.}^2}{n_{.j.}} - \frac{X_{....}^2}{N}$
<i>Radi(Paci)</i>	10	$\sum_{ijk} \frac{X_{ijk}^2}{n_{ijk}} - \sum_j \frac{X_{.j.}^2}{n_{.j.}}$
<i>Medix Paci</i>	9	$\sum_{ij} \frac{X_{ij.}^2}{n_{ij.}} - \sum_i \frac{X_{i...}^2}{n_{i...}} - \sum_j \frac{X_{.j.}^2}{n_{.j.}} + \frac{X_{....}^2}{N}$
<i>Medix Ra(P)</i>	10	$\sum_{ik} \frac{X_{i.k.}^2}{n_{i.k.}} - \sum_{ij} \frac{X_{ij.}^2}{n_{ij.}} - \sum_{ijk} \frac{X_{ijk}^2}{n_{ijk}} + \sum_j \frac{X_{.j.}^2}{n_{.j.}}$
<i>Error</i>	40	$\sum_{ijkl} X_{ijkl}^2 - \sum_{ik} \frac{X_{i.k.}^2}{n_{i.k.}}$
<i>total</i>	79	$\sum_{ijkl} X_{ijkl}^2 - \frac{X_{....}^2}{N}$

Y las pruebas *F* serían

<i>FV</i>	<i>ECM</i>	<i>F<sub>o</sub></i>
<i>Médico</i>	$\sigma^2 + \sigma_{MR}^2 + 4\sigma_{MP}^2 + 40\sigma_M^2$	$\frac{CM_M}{CM_{MP}}$
<i>Paciente</i>	$\sigma^2 + 4\sigma_R^2 + 8\sigma_P^2$	$\frac{CM_P}{CM_{R(P)}}$
<i>Radi(Paci)</i>	$\sigma^2 + 4\sigma_R^2$	$\frac{CM_{R(P)}}{CM_{Error}}$
<i>Medi xPaci</i>	$\sigma^2 + 2\sigma_{MR}^2 + 4\sigma_{MP}^2$	$\frac{CM_{MP}}{CM_{MR(P)}}$
<i>Médix Ra(P)</i>	$\sigma^2 + 2\sigma_{MR}^2$	$\frac{CM_{MR(P)}}{CM_{Error}}$
<i>Error</i>	$\sigma^2$	
<i>Total</i>		



Esto es, para probar el problema de partida, si los médicos miden lo mismo o no, la  $F$  es la división del cuadrado medio debido a médico por el cuadrado medio de la interacción médico/paciente.

**Prueba  $F$  aproximada.-**

Se puede dar el caso de que al observar las estimas de los  $CM$  se encuentre que un factor no tiene término con el que contrastarlo para hacer la prueba  $F$ , bien porque no haya ningún otro término que estime el mismo polinomio que el factor que se quiere contrastar, menos la varianza debida a dicho factor, o bien porque aunque lo haya, los coeficientes no son los mismos. En estos casos se hace una prueba  $F$  aproximada por el método de *Satterthwaite* del modo que veremos ahora. Por ejemplo, sea el modelo

$$X_{ijkl} = \mu + A_i + B_{j(i)} + c_k + AC_{ik} + BC_{kj(i)} + e_{ijkl}$$

siendo

- A antibiótico, y por tanto un factor de efectos fijos.
- B animales en los que se ha utilizado cada antibiótico, por lo tanto, un factor de efectos aleatorios.
- C días, tomados al azar, en los que se realizaron las mediciones, por lo tanto, un factor de efectos aleatorios.

Las estimas de los cuadrados medios son

	2	10	2	2	
	<b>F</b>	<b>A</b>	<b>A</b>	<b>A</b>	
	<i>i</i>	<i>j</i>	<i>k</i>	<i>l</i>	ECM
$e_{ijkl}$	1	1	1	1	$\sigma^2$
$BC_{ik(j)}$	1	1	1	1	$\sigma^2 + \sigma_{BC}^2$
$AC_{ik}$	0	b	1	1	$\sigma^2 + \sigma_{BC}^2 + b\sigma_{AC}^2$
$C_k$	a	b	1	1	$\sigma^2 + \sigma_{BC}^2 + ab\sigma_C^2$
$B_{j(i)}$	1	1	c	1	$\sigma^2 + \sigma_{BC}^2 + c\sigma_B^2$
$A_i$	0	b	c	1	$\sigma^2 + \sigma_{BC}^2 + b\sigma_{AC}^2 + c\sigma_B^2 + b\sigma_A^2$

En este caso las fuentes de variación  $B$ ,  $C$ , y  $AC$  se prueban con el cuadrado medio de  $BC$  y esta con el error, pero para probar el factor  $A$  no se tiene un denominador adecuado.

Si se realizara la prueba

$$F = \frac{CM_A}{CM_{AC}}$$

y diera no significativo al nivel de significación adecuado, se podría afirmar que ni  $\sigma^2_B$  ni  $\sigma^2_A$  tienen valores apreciables, por lo tanto serían no significativos ambos factores. Pero si esta  $F$  diera significativa no se podría distinguir a cuál de los dos factores ( $A$  o  $B$ ) es debida la significación, a no ser que la  $F = CM_B/CM_{BC}$  diera un valor menor o próximo a la unidad, lo cual significaría que  $\sigma^2_B=0$ .

Este problema de contrastar correctamente el factor  $A$  se resuelve mediante la  $F$  aproximada en forma de un cociente cuyo numerador y denominador sean combinaciones lineales de los  $CM$  de que disponemos en el análisis y de manera que la  $ECM$  del numerador contenga una varianza más que la del denominador, y que esta varianza sea la del factor que queremos probar. Además, las varianzas comunes deberán tener el mismo coeficiente.

También es aconsejable, que estas combinaciones lineales de los  $CM$  se hagan, siempre que se pueda, mediante sumas y no con restas, pues para un cierto error de las estimas del  $CM$  siempre repercutiría más en una cifra menor (resta) que en una cifra mayor (suma). Por otra parte, con números pequeños en el numerador y denominador, hay mayor peligro de que suba excesivamente el valor de  $F$ , es decir, la  $F$  tiende a la región de aceptación con números altos.

Los grados de libertad correspondientes a esta prueba serán el resultado de formar igual combinación lineal que la realizada para los  $CM$ .

En el ejemplo del modelo que se esta viendo se tendría que para probar el factor  $A$  la  $F$  adecuada sería

$$F_o = \frac{a_1 CM_A + a_2 CM_{BC}}{a_3 CM_B + a_4 CM_{AC}}$$

y los grados de libertad correspondientes serían

$$g_{\text{numerador}} = \frac{(a_1 CM_A + a_2 CM_{BC})^2}{\frac{(a_1 CM_A)^2}{g_A} + \frac{(a_2 CM_{BC})^2}{g_{BC}}}$$

$$g_{\text{denominador}} = \frac{(a_3 CM_B + a_4 CM_{AC})^2}{\frac{(a_3 CM_B)^2}{g_B} + \frac{(a_4 CM_{AC})^2}{g_{AC}}}$$

Estas cifras se redondearán para consultar la tabla  $F$  (Tabla 3).

Por ejemplo, supóngase que el análisis de varianza correspondiente al modelo anterior es

FV	gl	CM	ECM
Antibiótico (A)	4	30000	$\sigma^2 + \sigma_{BC}^2 + b\sigma_{AC}^2 + c\sigma_B^2 + bc\sigma_A^2$
Animal (B(A))	20	2800	$\sigma^2 + \sigma_{BC}^2 + c\sigma_B^2$
Día (C)	5	4000	$\sigma^2 + \sigma_{BC}^2 + ab\sigma_C^2$
Anti x Día	20	1500	$\sigma^2 + \sigma_{BC}^2 + b\sigma_{AC}^2$
Anim x Día	100	300	$\sigma^2 + \sigma_{BC}^2$
Error	150	200	$\sigma^2$

Como se dijo, la prueba  $F$  aproximada para el factor A es

$$F_o = \frac{30000 + 300}{2800 + 1500} = 7.046$$

ya que  $a_1 = a_2 = a_3 = a_4$ , los grados de libertad son

$$gl_{\text{numerador}} = \frac{(30000 + 300)^2}{\frac{30000^2}{4} + \frac{300^2}{100}} = 4.1$$

$$gl_{\text{denominador}} = \frac{(2800 + 1500)^2}{\frac{2800^2}{20} + \frac{1500^2}{20}} = 36.65$$

Se contrasta el valor de  $F_o = 7.04$  con el valor  $F_{(4,37; 0.05)} = 3.87$ , por lo que es significativa la  $F_o$ , y se concluye que los antibióticos son diferentes.

### Ejemplo.-

Recuérdese el ejemplo que se vio en el epígrafe *Generalización del método para el análisis de modelos factoriales con repetición con más de dos factores*, en el que se planteaba las incidencias que tienen dos dietas en la producción de leche en cabras y se tenían en cuenta, también, el número de lactación de la cabra y la época de parto de la cabra. Esta experiencia responde a un modelo trifactorial en el que todos los factores son de efectos fijos, por lo que en todas las pruebas de hipótesis el término de contraste es el cuadrado medio del error.

Supóngase los mismos datos de aquel ejemplo, pero los factores son: dos *tratamientos* (las mismas dos dietas), tres *rebaños* e individuos procedentes de dos *líneas*, una seleccionada y otra sin seleccionar. Los datos son

$T_1$				$T_2$					
	$R_1$	$R_2$	$R_3$		$R_1$	$R_2$	$R_3$		
$L_1$	325	374	364	$L_1$	375	401	392		
	345	395	359		384	395	385		
$L_2$	304	352	336	$L_2$	352	395	362		
	310	348	328		349	360	365		
4140			2744			3020		2891	4515

Factores	Símbolo	Tipo	Relación	Niveles
Tratamiento	(T)	Fijo	↕	$i=2$
Rebaño	(R)	Aleatorio	↕	$j=3$
Línea	(L)	Aleatorio	↕	$k=2$
Repetición	(e)	Aleatorio	↑	$l=2$

Por lo que el modelo lineal es

$$X_{ijkl} = \mu + T_i + R_j + L_k + TR_{ij} + TL_{ik} + RL_{jk} + TRL_{ijk} + e_{ijkl}$$

Las estimas de los cuadrados medios son

	2	3	2	2	
	F	A	A	A	ECM
	i	j	k	l	
$e_{ijkl}$	1	1	1	1	$\sigma^2$
$TRL_{ijk}$	0	1	1	2	$\sigma^2 + 2\sigma_{TRL}^2$
$RL_{ik}$	2	1	1	2	$\sigma^2 + 4\sigma_{RL}^2$
$TL_{ik}$	0	3	1	2	$\sigma^2 + 2\sigma_{TRL}^2 + 6\sigma_{TL}^2$
$TR_{ij}$	0	1	2	2	$\sigma^2 + 2\sigma_{TRL}^2 + 4\sigma_{TR}^2$
$L_k$	2	3	1	2	$\sigma^2 + 4\sigma_{RL}^2 + 12\sigma_L^2$
$R_j$	2	1	2	2	$\sigma^2 + 4\sigma_{RL}^2 + 8\sigma_R^2$
$T_i$	0	3	2	2	$\sigma^2 + 4\sigma_{TR}^2 + 6\sigma_{TL}^2 + 2\sigma_{TRL}^2 + 12\sigma_T^2$

Las pruebas F serían

<i>FV</i>	<i>ECM</i>	<i>F<sub>o</sub></i>
<i>Trata</i>	$\sigma^2 + 4\sigma_{TR}^2 + 6\sigma_{TL}^2 + 2\sigma_{TRL}^2 + 12\sigma_T^2$	$\frac{CM_T}{CM_{TR} + CM_{TL} - CM_{TRL}}$
<i>Rebaño</i>	$\sigma^2 + 4\sigma_{RL}^2 + 8\sigma_R^2$	$\frac{CM_R}{CM_{RL}}$
<i>Línea</i>	$\sigma^2 + 4\sigma_{RL}^2 + 12\sigma_L^2$	$\frac{CM_L}{CM_{RL}}$
<i>Trax Reb</i>	$\sigma^2 + 2\sigma_{TRL}^2 + 4\sigma_{TR}^2$	$\frac{CM_{TR}}{CM_{TRL}}$
<i>Tra xLín</i>	$\sigma^2 + 2\sigma_{TRL}^2 + 6\sigma_{TL}^2$	$\frac{CM_{TL}}{CM_{TRL}}$
<i>Rebx Lín</i>	$\sigma^2 + 4\sigma_{RL}^2$	$\frac{CM_{RL}}{CM_{Error}}$
<i>TraxRebxLí n</i>	$\sigma^2 + 2\sigma_{TRL}^2$	$\frac{CM_{TRL}}{CM_{Error}}$
<i>Error</i>	$\sigma^2$	
<i>total</i>		

Como se ve, la *F* para probar el factor *Tratamientos* es una *F* aproximada, pues ninguna de las estimas de los cuadrados medios de las demás fuentes de variación estiman lo mismo que el cuadrado medio de tratamientos menos la varianza debida a los tratamientos. Sin embargo, el denominador de esta *F* aproximada estima lo mismo que el numerador menos la varianza de los tratamientos, por lo que cumple las condiciones deseadas.

Los grados de libertad de esta *F* son, para el numerador los de los tratamientos, es decir, uno; y para el denominador

$$g_{\text{denominador}} = \frac{(CM_{TR} + CM_{TL} - CM_{TRL})^2}{\frac{CM_{TR}^2}{g_{TR}} + \frac{CM_{TL}^2}{g_{TL}} - \frac{CM_{TRL}^2}{g_{TRL}}}$$

$$g_{\text{denominador}} = \frac{(282.2 + 51 - 28.75)^2}{\frac{282.2^2}{2} + \frac{51^2}{1} - \frac{28.75^2}{2}} = 2.16$$

Las *F* de contraste son, para *T*, *L* y *TL*,  $F_{(1,2; 0.05)} = 18.5$ , para *R* y *TR*,  $F_{(2,2; 0.05)} = 19.0$ , y para *RL* y *TRL*  $F_{(2,12; 0.05)} = 3.89$

Por lo que se concluye que los efectos de los tres factores principales son significativos y ninguna interacción influye en la variabilidad de los datos.

**Archivo del programa SAS (C9-2.SAS).-**

```

title 'Modelo trifactorial mixto (fijo el primer factor)';
option ps=65 ls=75;
data fact3;
infile 'c9-2.dat';
input trata $ rebano $ linea $ prod @@;
proc glm;
  class trata rebano linea;
  model prod = trata rebano linea trata*rebano trata*linea
            rebano*linea trata*rebano*linea;
run;
title 'Pruebas para TRATA y todas sus interacciones';
random rebano linea trata*rebano trata*rebano trata*linea rebano*linea
       trata*rebano trata*rebano*linea / test;
run;
title 'Pruebas para REBAÑO y LINEA, y sus interacciones';
random rebano linea rebano*linea / test;
run;

```

**Archivo de datos (C9-2.DAT).-**

T1 R1 L1 325	T1 R2 L1 374	T1 R3 L1 364
T1 R1 L1 345	T1 R2 L1 395	T1 R3 L1 359
T1 R1 L2 304	T1 R2 L2 352	T1 R3 L2 336
T1 R1 L2 310	T1 R2 L2 348	T1 R3 L2 328
T2 R1 L1 375	T2 R2 L1 401	T2 R3 L1 392
T2 R1 L1 384	T2 R2 L1 395	T2 R3 L1 385
T2 R1 L2 352	T2 R2 L2 395	T2 R3 L2 362
T2 R1 L2 349	T2 R2 L2 360	T2 R3 L2 365

**Archivo de resultados (C9-2.LST).-**

Dependent Variable: PROD						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	11	15922.1250	1447.4659	14.53	0.0001	
Error	12	1195.5000	99.6250			
Corrected Total	23	17117.6250				
Source	R-Square	C.V.	Root MSE	PROD Mean		
	0.930160	2.767759	9.98123	360.625		
Source	DF	Type III SS	Mean Square	F Value	Pr > F	
TRATA	1	5859.37500	5859.37500	58.81	0.0001	
REBANO	2	4767.75000	2383.87500	23.93	0.0001	
LINEA	1	4620.37500	4620.37500	46.38	0.0001	
TRATA*REBANO	2	564.25000	282.12500	2.83	0.0983	
TRATA*LINEA	1	51.04167	51.04167	0.51	0.4878	
REBANO*LINEA	2	1.75000	0.87500	0.01	0.9913	
TRATA*REBANO*LINEA	2	57.58333	28.79167	0.29	0.7541	

TRATA	Var(Error) + 2 Var(TRATA*REBANO*LINEA)					
	+ 6 Var(TRATA*LINEA) + 4 Var(TRATA*REBANO) + Q(TRATA)					
REBANO	Var(Error) + 2 Var(TRATA*REBANO*LINEA)					
	+ 4 Var(REBANO*LINEA) + 4 Var(TRATA*REBANO)					
	+ 8 Var(REBANO)					
LINEA	Var(Error) + 2 Var(TRATA*REBANO*LINEA)					
	+ 4 Var(REBANO*LINEA) + 6 Var(TRATA*LINEA)					
	+ 12 Var(LINEA)					
TRATA*REBANO	Var(Error) + 2 Var(TRATA*REBANO*LINEA)					
	+ 4 Var(TRATA*REBANO)					
TRATA*LINEA	Var(Error) + 2 Var(TRATA*REBANO*LINEA)					
	+ 6 Var(TRATA*LINEA)					
REBANO*LINEA	Var(Error) + 2 Var(TRATA*REBANO*LINEA)					
	+ 4 Var(REBANO*LINEA)					
TRATA*REBANO*LINEA	Var(Error) + 2 Var(TRATA*REBANO*LINEA)					
Source: TRATA						
Error: MS (TRATA*REBANO) + MS (TRATA*LINEA) - MS (TRATA*REBANO*LINEA)						
	DF	Type III MS	Denominator	Denominator	F Value	Pr > F
	1	5859.375	DF	MS	19.2505	0.0417
			2.16	304.375		
Source: TRATA*REBANO						
Error: MS (TRATA*REBANO*LINEA)						
	DF	Type III MS	Denominator	Denominator	F Value	Pr > F
	2	282.125	DF	MS	9.7988	0.0926
			2	28.791666667		
Source: TRATA*LINEA						
Error: MS (TRATA*REBANO*LINEA)						
	DF	Type III MS	Denominator	Denominator	F Value	Pr > F
	1	51.041666667	DF	MS	1.7728	0.3145
			2	28.791666667		
Source: TRATA*REBANO*LINEA						
Error: MS (Error)						
	DF	Type III MS	Denominator	Denominator	F Value	Pr > F
	2	28.791666667	DF	MS	0.2890	0.7541
			12	99.625		
Source						
Type III Expected Mean Square						
TRATA	Var(Error)					
	+ Q (TRATA, TRATA*REBANO, TRATA*LINEA, TRATA*REBANO*LINEA)					
REBANO	Var(Error) + 4 Var(REBANO*LINEA) + 8 Var(REBANO)					
	+ Q (TRATA*REBANO, TRATA*REBANO*LINEA)					
LINEA	Var(Error) + 4 Var(REBANO*LINEA) + 12 Var(LINEA)					
	+ Q (TRATA*LINEA, TRATA*REBANO*LINEA)					
TRATA*REBANO	Var(Error) + Q (TRATA*REBANO, TRATA*REBANO*LINEA)					
TRATA*LINEA	Var(Error) + Q (TRATA*LINEA, TRATA*REBANO*LINEA)					
REBANO*LINEA	Var(Error) + 4 Var(REBANO*LINEA)					
	+ Q (TRATA*REBANO*LINEA)					
TRATA*REBANO*LINEA	Var(Error) + Q (TRATA*REBANO*LINEA)					
Source: REBANO *						
Error: MS (REBANO*LINEA)						
	DF	Type III MS	Denominator	Denominator	F Value	Pr > F
	2	2383.875	DF	MS	2724	0.0004
			2	0.875		
* - This test assumes one or more other fixed effects are zero.						

Source: LINEA *						
Error: MS(REBANO*LINEA)						
DF	Type III MS	Denominator	Denominator		F Value	Pr > F
1	4620.375	2	0.875		5280	0.0002
* - This test assumes one or more other fixed effects are zero.						
Source: REBANO*LINEA						
Error: MS(Error)						
DF	Type III MS	Denominator	Denominator		F Value	Pr > F
2	0.875	12	99.625		0.008783	0.9913
Source: TRATA*REBANO*LINEA						
Error: MS(Error)						
DF	Type III MS	Denominator	Denominator		F Value	Pr > F
2	28.791666667	12	99.625		0.2890	0.7541

### Estima conjunta (*englobe*) de la significación de varios factores en modelos compuestos.-

Cuando, en un análisis, la prueba  $F$  de un factor da un valor que estadísticamente pueda considerarse como la unidad, indica que los dos factores, el numerador y el denominador, están estimando lo mismo y por tanto no tiene sentido mantener ambos términos pudiendo englobarlos en uno sólo, obteniendo así una mejor estima del  $CM$  del factor que actuaba como denominador.

Si la  $F$  obtenida es menor o igual que la unidad se puede proceder a hacer el *englobe* directamente. Por el contrario, si resulta  $F_o > 1$  es ya más difícil e incluso discutible tomar tal decisión, por lo cual se debe tener una norma, más o menos flexible, en la que basarse. Esta norma podría consistir en tomar un nivel de  $\alpha$  mayor que el usado a lo largo del análisis y contrastar la  $F$  obtenida con la  $F$  correspondiente ese nivel, haciendo el *englobe* si resulta menor. Se toma un  $\alpha$  mayor con el fin de tener más seguridad de que los dos  $CM$  están estimando lo mismo.

El criterio normalmente usado es utilizar un nivel  $\alpha=0.25$  para esta prueba previa de significación, que indicará si se procede a realizar el *englobe*.

Lo lógico es que, cuando se presenten casos de  $F < 1$ , se empiece englobando las interacciones de orden más alto, a continuación las de orden inferior, etc. También se puede aplicar esta técnica a cualquier efecto principal en el que no se esté muy interesado pero que apareció en el modelo como clasificación obligada para tratar de reducir la magnitud del error. En este caso, al comprobar que no hay indicios de su influencia, se debe devolver su aparente variabilidad a su correspondiente término del error. Si se trata de un factor básico en nuestro estudio, no tiene sentido el hacer el *englobe*.

Veamos, en el ejemplo de dos factores con repetición, cuándo es aconsejable realizar el *englobe* o no.



<i>FV</i>	<i>gl</i>	<i>ECM</i>
<i>Tratamientos</i>	<i>t-1</i>	$\sigma^2 + n\sigma_{TR}^2 + m\sigma_T^2$
<i>Rebaño</i>	<i>r-1</i>	$\sigma^2 + tn\sigma_R^2$
<i>TxR</i>	$(t-1)(r-1)$	$\sigma^2 + n\sigma_{TR}^2$
<i>Error</i>	<i>N-tr</i>	$\sigma^2$
<i>total</i>	<i>N-1</i>	

Supóngase que se han realizado cinco experimentos en los que se han obtenido, respectivamente, los siguientes valores de *F* para la interacción

$$F_1 = 0.80; \quad F_2 = 1.40; \quad F_3 = 1.84; \quad F_4 = 4.52; \quad F_5 = 6.94$$

Supóngase que los experimentos son iguales al del ejemplo visto para modelos factoriales con repetición, esto es, la interacción tiene 2 y 6 grados de libertad, por lo que se pueden contrastar con las siguientes *F* de las tablas

$$F_{(2,6; 0.05)} = 5.14; \quad F_{(2,6; 0.10)} = 3.46; \quad F_{(2,6; 0.25)} = 1.76$$

En el primer caso se debe hacer el englobe, ya que al ser  $F_1 < 1$  se puede afirmar que  $\sigma_{RT}^2$  es prácticamente cero.

En el cuarto y quinto caso no se debe hacer el englobe pues la interacción es significativa al  $\alpha=0.10$  y  $\alpha=0.05$  respectivamente.

En el segundo caso se ve que  $F_o$  es algo mayor que la unidad, pero no es significativa ni con un nivel de significación de  $\alpha=0.25$  ( $F_2 < F_{(2,6; 0.25)}$ ) por lo tanto se podría hacer el englobe.

En el tercer caso, la  $F_o$  es relativamente pequeña, pero no es aconsejable realizar el englobe porque resulta significativa al  $\alpha=0.25$ .

La utilidad de englobar términos en uno sólo se basa en que al agrupar con el error las componentes que no tienen significación, la estima de dicho error va a ser mucho más precisa por estar basado en más datos. Cuando la  $F_o$  de la prueba del factor que se agrupa ha sido menor que la unidad el nuevo error disminuirá de valor, lo que resulta favorable en toda prueba *F*. Si la  $F_o > 1$ , el error aumentará ligeramente. En todo caso, en ambas situaciones aumentan los *gl* para la prueba y como consecuencia su sensibilidad.

También es preciso realizar el englobe cuando el objetivo del estudio es la estima de los componentes de la varianza, con objeto de emplearlos en la estima de otros parámetros, ya que en el caso de que la  $F_o$  sea menor que la unidad, dará estimas de varianza menores de cero, por lo que se tendrá que eliminar esas incongruencias.

Para realizar el englobe de dos o más factores se pondrá por *SC* la suma de las *SC* de los factores englobados, y los *gl* serán también la suma de los *gl* individuales. El nuevo *CM* será el resultado de dividir la *SC* resultante por los *gl* resultante.

### Ejemplo.-

En una prueba con gallinas alojadas en el suelo se estudian cuatro tratamientos. Como se dispone de 12 lotes experimentales se hacen grupos de tres repeticiones a los que se les aplica, completamente al azar, cada tratamiento. El carácter estudiado es peso del huevo y como este varía con la edad de la gallina, se realiza las pesadas en cinco edades o periodos de las gallinas.

El modelo es

$$X_{ijkl} = \mu + T_i + R_{j(i)} + P_k + TP_{ik} + RP_{j(i)k} + e_{ijkl}$$

siendo las repeticiones un factor de efectos aleatorios, y tratamientos y periodos factores de efectos fijos.

Una vez realizadas las operaciones, se tiene el siguiente análisis

<i>FV</i>	<i>gl</i>	<i>SC</i>	<i>CM</i>	<i>ECM</i>	<i>F<sub>o</sub></i>
<i>Tratamiento</i>	3	42.62	14.20	$\sigma^2 + 5\sigma_R^2 + 15\sigma_T^2$	1.37ns
<i>Repetición</i>	8	82.84	10.35	$\sigma^2 + 5\sigma_R^2$	0.60ns
<i>Periodo</i>	4	5699.72	1424.93	$\sigma^2 + \sigma_{RP}^2 + 12\sigma_P^2$	83.91**
<i>TraxPer</i>	12	452.70	37.72	$\sigma^2 + \sigma_{RP}^2 + 3\sigma_{TP}^2$	2.22*
<i>RexPe+Error</i>	32	543.57	16.98	$\sigma^2 + \sigma_{RP}^2$	

Por experiencias anteriores se sabe que la interacción *repeticiones/periodo* no existe, por tanto  $\sigma_{RP}^2 = 0$ . En este caso, la prueba del factor *repeticiones* será con *RP+Error*, que estima solo el error. Una vez realizada la prueba *F* se ve que dá menor que la unidad, lo que indica que *R* y *RP* están estimando lo mismo, debiéndose al azar el que el valor de *F<sub>o</sub>* sea menor que la unidad. Este es un ejemplo en el que debe hacerse el englobe de *R* y su término de contraste, con lo cual el error, en lugar de tener 32 *gl*, tendrá 40 y el nuevo *CM* será 15.66.

Por otra parte, aunque la *F<sub>o</sub>* de tratamientos ha dado muy baja, no es lógico hacer el englobe de este factor, ya que es uno de los objetivos principales del análisis.

El análisis quedaría

<i>FV</i>	<i>gl</i>	<i>SC</i>	<i>CM</i>	<i>ECM</i>	<i>F<sub>o</sub></i>
<i>Tratamiento</i>	3	42.62	14.20	$\sigma^2 + 5\sigma_R^2 + 15\sigma_T^2$	0.91ns
<i>Periodo</i>	4	5699.72	1424.93	$\sigma^2 + \sigma_{RP}^2 + 12\sigma_P^2$	91.05**
<i>Tra Per</i>	12	452.70	37.72	$\sigma^2 + \sigma_{RP}^2 + 3\sigma_{TP}^2$	2.41*
<i>RexPe+Pe+Error</i>	40	626.41	15.66	$\sigma^2 + \sigma_{RP}^2$	

El factor *Periodo* ha sido muy significativo, como era de esperar, ya que la producción varía enormemente con la edad de las gallinas. El que la interacción *TP* resulte significativa sin que lo sean los tratamientos nos debe poner en guardia para su interpretación pues esto podría ser un artefacto matemático debido a la alta significación de los periodos, los cuales han tenido tanta influencia, que pequeñas variaciones producidas por los tratamientos traen consigo variaciones grandes al estudiarlos por periodos separados. No parece lógico pensar, en este caso, que los efectos de los tratamientos tengan una relación diferente en diferentes periodos, esto es, que unos tratamientos tengan efectos positivos en uno, o unos, periodo y efectos negativos en otro, u otros, periodo de manera que se anulen los efectos de unos periodos con los de otros periodos.

En otro estudio, los tratamientos se aplicaron a gallinas alojadas en baterías y los resultados del análisis fueron los siguientes

<i>FV</i>	<i>gl</i>	<i>SC</i>	<i>CM</i>	<i>ECM</i>	<i>F<sub>o</sub></i>
<i>Tratamiento</i>	3	867.87	289.29	$\sigma^2 + 5\sigma_R^2 + 15\sigma_T^2$	9.331**
<i>Repetición</i>	8	247.92	30.99	$\sigma^2 + 5\sigma_R^2$	1.25ns
<i>Periodo</i>	4	7161.28	1790.32	$\sigma^2 + \sigma_{RP}^2 + 12\sigma_P^2$	72.51**
<i>TraxPer</i>	12	139.68	11.64	$\sigma^2 + \sigma_{RP}^2 + 3\sigma_{TP}^2$	0.47ns
<i>RexPe+Error</i>	32	790.08	24.69	$\sigma^2 + \sigma_{RP}^2$	

En este caso está claro que hay que hacer el englobe de *TP* y *RP* puesto que, parece lógico que  $\sigma_{TP}^2$  no exista ( $F < 0$ ). Ahora bien ¿Tendríamos que incluir en el englobe la componente *repetición*? El criterio apuntado anteriormente es que se pruebe con el nivel  $\alpha=0.25$  para ver si realmente sigue siendo ns. El valor de *F* a este nivel de significación es  $F_{(1,12; 0,25)}=1.36$ . Como es no significativa al  $\alpha=0.25$ , la mejor estima del error vendrá dada por el *CM* conjunto de  $(RP)+(TP)+R$ .

Como resumen se puede decir que, en principio, hay que hacer la clasificación más completa posible incluyendo en ella todos los factores que puedan tener influencia, y una vez halladas las significaciones ir eliminando aquellas componentes que no sean importantes mediante el englobe, para tener así una mejor estima del error (o de otro término que actúe como error) y por ello una mayor exactitud en la

prueba  $F$  al aumentar el número de grados de libertad. Además esta estima del error será, como se verá más adelante, la que se utilizará como base para la comparación de medias cuando se tiene factores de efectos fijos.

### **Bibliografía:**

- Cochran, WG y Cox, GM.* 1965. DISEÑOS EXPERIMENTALES. Ed. TRILLAS. México.
- Li, Ching Chun.* 1969. INTRODUCCIÓN A LA ESTADÍSTICA EXPERIMENTAL. Ed. Omega. Barcelona.
- Lite, TM, y Jackson Hills, F.* 1987. MÉTODOS ESTADÍSTICOS PARA LA INVESTIGACIÓN EN LA AGRICULTURA. Ed. TRILLAS. México.
- Littel, R.C., Freund, R.J. and Spector, P.C.* 1991. SAS® FOR LINEAR MODELS. SAS Institute Inc., Cary, NC, USA.
- Ostle, B.* 1965. ESTADÍSTICA APLICADA. Ed. Limusa-Wiley. México.
- Ruiz-Maya, L.* 1983. MÉTODOS ESTADÍSTICOS DE INVESTIGACIÓN. Ed. INE. Madrid.
- Sokal, R.R. y Rohlf, F.J.* 1994. BIOMETRY. Ed. W.H.FREEMAN. San Francisco.
- Snedecor, G.W. y Cochran, W.G.* 1971. MÉTODOS ESTADÍSTICOS. Ed. C.E.C.S.A. México.
- Steel, R.* 1996. PRINCIPLES AND PROCEDURES OF STATISTICS. Ed. McGRAW-HILL Education. New York .
- SAS Institute Inc.* 1990. SAS/STAT® USER'S GUIDE. Volume 1 and 2. Cary, NC, USA.



## **CAPÍTULO 10**

# **Comparación de Medias**



## Comparación de Medias

Se ha dicho anteriormente que si el factor (o factores) es de efectos fijos, lo lógico es que se quiera saber qué media, o medias, es diferente de otras medias o media; o cuáles grupos de medias son diferentes de otros grupos de medias o de medias aisladas; o cuál es la tendencia que siguen los efectos de los diferentes tratamientos (es decir, las medias) o niveles del factor si estos son de tipo cuantitativo.

Hay, por tanto, tres planteamientos para las comparaciones de medias

- a) Comparaciones planeadas o a priori
- b) Comparaciones múltiples de medias o a posteriori.
- c) Contrastes ortogonales.

Se tiene comparaciones planeadas cuando las pruebas se planifican antes del experimento y el análisis se realiza independientemente de la prueba del factor completo, como consecuencia de los planteamientos y estructura de dicho experimento. Por ejemplo, podría ser de interés el comparar si los diferentes tratamientos, tomados en conjunto, son diferentes del control, aun cuando la prueba  $F$  para tratamientos hubiera sido o no significativa.

Por el contrario, se tiene comparaciones múltiples, no planeadas, cuando, una vez realizado el experimento y el análisis de varianza y obtenida una  $F$  significativa, se quiere saber en qué sentido se producen dichas diferencias entre las medias, es decir, cuál o cuáles tratamientos tienen una media mayor y cuál o cuáles tratamientos tienen una media menor, es decir, es una prueba sugerida por los datos o de razonamiento posterior a los hechos. Como se ve claramente, se haría una comparación múltiple sólo si el ANOVA ha sido significativo y una vez vistos los datos, pues en caso contrario no tienen sentido, de ahí los nombres que reciben.

Se tiene un contraste ortogonal cuando se tienen factores cuantitativos con niveles igualmente espaciados, y se quiere saber cual es la tendencia que siguen estos niveles. Y sobre todo, lo que se querrá en muchas ocasiones es estimar la respuesta óptima (experimentos de *superficie respuesta*).



## Comparaciones o pruebas planeadas.-

Hay dos maneras de realizar estas pruebas a priori. Se puede subdividir la suma de los cuadrados del tratamiento y los grados de libertad en comparaciones separadas, realizando éstas como parte del análisis de la varianza, utilizando el estadístico de prueba  $F$ . Estas pruebas son sencillas y enlazan perfectamente con el desarrollo completo del ANOVA.

La regla general para realizar una prueba planeada es muy simple, básicamente es la misma regla que se dio anteriormente para la obtención de la suma de cuadrados para cualquier conjunto de grupos. Para comparar  $k$  grupos (subconjunto de los  $t$  tratamientos) de tamaño  $n_i$  se toma la suma de cada grupo, se eleva al cuadrado y se divide por el tamaño muestral  $n_i$ , sumando, posteriormente, los  $k$  cocientes de esta manera obtenidos. A esta suma se le restará un término de corrección, el cual será el cuadrado de la suma de todos los grupos que se comparan dividido por el número de datos de la suma total. Si la comparación implica a todos los grupos del ANOVA, el término de corrección será el ya conocido ( $TC$ ). Y si la comparación implica únicamente a algunos grupos del ANOVA, el término de corrección será diferente, estando restringido únicamente a estos grupos. Más adelante se dará la regla para estimar estas sumas de cuadrados en base a unos coeficientes, de manera que llevan implícito la substracción del término de corrección.

Estas reglas se entenderán mejor con un ejemplo.

Supongamos que se están analizando cuatro tipos de piensos con  $n$  observaciones en cada uno. Con arreglo al esquema

$T_1$	$T_2$	$T_3$	$T_4$
$n$	$n$	$n$	$n$

El análisis tendría las fuentes de variación y los grados de libertad:

$FV$	$gl$
Tratamientos	3
Error	$4(n-1)$

Si estos piensos tienen una composición determinada en proteínas y niveles de energía, tal como

	Proteínas	Energía
Pienso 1	20%	3000
Pienso 2	20%	3500
Pienso 3	22%	3000
Pienso 4	22%	3500

Recordando lo que se estudió en los epígrafes anteriores referidos a los experimentos factoriales, se podría tener este otro esquema de la misma experiencia

	$E_1$	$E_2$
$P_1$	$T_1$	$T_2$
$P_2$	$T_3$	$T_4$

como consecuencia de que antes de realizar la experiencia se decidió examinar las siguientes comparaciones

- 1) Comparar la proteína alta con la baja.
- 2) Comparar la energía alta con la baja.
- 3) Comparar la proteína baja/energía baja y proteína alta/energía alta con la proteína baja/energía alta y proteína alta/energía baja.

Esto significa que se quiere, además, hacer las siguientes pruebas de hipótesis

- 1)  $H_0: -\mu_1 - \mu_2 + \mu_3 + \mu_4 = 0$
- 2)  $H_0: -\mu_1 + \mu_2 - \mu_3 + \mu_4 = 0$
- 3)  $H_0: +\mu_1 - \mu_2 - \mu_3 + \mu_4 = 0$

Por tanto, el análisis en este caso podría ser

<i>FV</i>	<i>gl</i>
<i>Tratamientos</i>	3
<i>Calorías</i>	1
<i>Proteínas</i>	1
<i>Cal x Pro</i>	1
<i>Error</i>	4( <i>n</i> -1)

Como se ve, es el mismo análisis que el anterior, en el que se ha descompuesto los efectos encontrados para los piensos o tratamientos en los tres componentes más lógicos, dada la composición de los mismos.

Los coeficientes para la obtención de los nuevos contrastes, dadas las hipótesis planteadas más arriba, son

	$C_1$	$C_2$	$C_3$	$C_4$	$\Sigma C_i$
1) <i>Calorías</i>	-1	1	-1	1	0
2) <i>Proteínas</i>	-1	-1	1	1	0
3) <i>Cal Pro</i>	1	-1	-1	1	0

Por tanto, las pruebas planeadas o a priori, se pueden formar con los diferentes niveles de un tratamiento, asignándole a cada nivel un coeficiente de forma que se cumplan las condiciones de ortogonalidad que son:

- 1) Dentro de una misma comparación, la suma de coeficientes ha de ser cero.
- 2) Entre los varios contrastes que se puedan formar dentro de un mismo factor, la suma de productos ordenados de los coeficientes ha de ser nula, tomando todas las comparaciones dos a dos.

Al cumplir las condiciones de ortogonalidad, estos contrastes son independientes.

En esencia, se trata de descomponer la suma de cuadrados correspondientes al factor considerado en otras sumas parciales que divididas por los  $g/$  darán un nuevo  $CM$  para cada contraste. La prueba  $F$  se hará siempre con el término de error utilizado para probar el factor completo.

No es necesario agotar todos los  $g/$ , es decir, hacer tantos contrastes como grados de libertad tenga la fuente de variación descompuesta, pues si hay alguno que complete la descomposición ortogonal y no se está interesado en él, no sirve de nada el continuar, pudiéndolo dejar para una  $SC$  residual. Con la suma de cuadrados residual, dividida por los grados de libertad residuales, se puede comprobar si los contrastes analizados agotaron ya toda la variabilidad atribuida al factor o aún queda algo de ella en lo que no se ha descompuesto.

Por otra parte, puede darse el caso de estar interesados en más contrastes de los que se pueden hacer con la condición de ortogonalidad y el límite de los  $g/$  totales de la fuente de variación descompuesta. En este caso se podría hacer más de una descomposición de las  $SC$ , incluyendo en una unos contrastes y en otra (u otras) los restantes en que se esté interesado. En este caso hay que tener cuidado en la interpretación de las pruebas  $F$  puesto que si dentro de cada descomposición los contrastes son independientes, puede que no lo sean al considerar los de varias descomposiciones.

Veamos si en la descomposición anterior se cumplen las condiciones de ortogonalidad

<i>Producto de coeficientes</i>									
1	×	2	=	+1	-1	-1	+1	=	0
1	×	3	=	-1	-1	+1	+1	=	0
2	×	4	=	-1	+1	-1	+1	=	0

<i>Suma de coeficientes</i>							
1	=	-1	+1	-1	+1	=	0
2	=	-1	-1	+1	+1	=	0
3	=	+1	-1	-1	+1	=	0

Como se ve, estos contrastes cumplen las condiciones de ortogonalidad: suma de coeficientes igual a cero y suma de productos, dos a dos, también igual a cero.

Aunque estos coeficientes se comprenden intuitivamente, tienen una justificación matemática, que se estudia a continuación

Si

$$\bar{X}_1, \bar{X}_2, \bar{X}_3, \text{ y } \bar{X}_4.$$

son las medias de cada tratamiento, las medias para cada uno de los dos niveles de calorías y proteínas son

$$\bar{C}_1 = \frac{\bar{X}_1 + \bar{X}_3}{2}; \quad \bar{P}_1 = \frac{\bar{X}_1 + \bar{X}_3}{2}$$

$$\bar{C}_2 = \frac{\bar{X}_2 + \bar{X}_4}{2}; \quad \bar{P}_2 = \frac{\bar{X}_3 + \bar{X}_4}{2}$$

Las diferencias de medias entre los dos niveles de calorías o proteínas son

$$D_{\bar{C}} = \bar{C}_2 - \bar{C}_1 = \frac{c_1 \bar{X}_1 + c_2 \bar{X}_2 + c_3 \bar{X}_3 + c_4 \bar{X}_4}{2} = \frac{-\bar{X}_1 + \bar{X}_2 - \bar{X}_3 + \bar{X}_4}{2}$$

$$D_{\bar{P}} = \bar{P}_2 - \bar{P}_1 = \frac{c_1 \bar{X}_1 + c_2 \bar{X}_2 + c_3 \bar{X}_3 + c_4 \bar{X}_4}{2} = \frac{-\bar{X}_1 - \bar{X}_2 + \bar{X}_3 + \bar{X}_4}{2}$$

En el caso de la interacción, se tendrá que comparar los niveles en los que las calorías y proteínas figuran, las dos, en cantidad alta y cantidad baja, es decir, en la *misma cantidad (MC)* con los niveles en que están a *distinta cantidad (DC)*, esto es

$$\overline{MC} = \frac{\bar{X}_1 + \bar{X}_4}{2}$$

$$\overline{DC} = \frac{\bar{X}_2 + \bar{X}_3}{2}$$

La diferencia es

$$D = \overline{MC} - \overline{DC} = \frac{c_1 \bar{X}_1 + c_2 \bar{X}_2 + c_3 \bar{X}_3 + c_4 \bar{X}_4}{2} = \frac{+\bar{X}_1 - \bar{X}_2 - \bar{X}_3 + \bar{X}_4}{2}$$

Por tanto, se pueden dar las siguientes reglas para la determinación de los coeficientes.

- 1) Si se van a comparar dos grupos de igual tamaño, simplemente se asignan coeficientes +1 a los miembros de un grupo y -1 a los integrantes del otro grupo. No importa a qué grupo se le asigne los coeficientes positivos o negativos.
- 2) En la comparación de grupos que contienen distintos números de tratamientos, asígnese al primer grupo tantos coeficientes como número de tratamientos tenga el segundo grupo; y a este último, tantos coeficientes, del signo opuesto, como número de tratamientos tenga el

primer grupo. Por ejemplo, si entre cinco tratamientos se quiere comparar los dos primeros con los tres últimos, los coeficientes serían +3, +3, -2, -2, -2.

- 3) Redúzcanse los coeficientes a los enteros más pequeños posibles. Por ejemplo, en la comparación de un grupo de dos tratamientos con un grupo de cuatro se tendrá (regla segunda) los coeficientes +4, +4, -2, -2, -2, -2 pero éstos pueden reducirse, dividiendo por dos, a los coeficientes +2, +2, -1, -1, -1, -1.
- 4) Los coeficientes de la interacción siempre pueden determinarse mediante la multiplicación de los coeficientes correspondientes de los efectos principales.

Otro ejemplo puede ser: en los pollos de carne, las hembras y los machos no tienen el mismo crecimiento. Si se quiere analizar si deben criarse por sexos separados o en conjunto, se pueden considerar tres tratamientos

$T_1$	$T_2$	$T_3$
hembras	machos	mezcla
$n$	$n$	$n$

Y el análisis sería

<i>FV</i>	<i>gl</i>
Tratamientos	2
Error	$3(n-1)$

Suponiendo que los tratamientos han dado significativos, se puede hacer dos comparaciones ya que se tiene dos *gl*, los más lógicos serían: machos vs hembras y crianza juntos vs crianza separados.

	Hembras	Machos	Mezcla
$C_1$ Machos vs Hembras	1	-1	0
$C_2$ Juntos vs Separados	1	1	-2

El análisis quedaría

<i>FV</i>	<i>gl</i>
Tratamientos	2
$C_1$	1
$C_2$	1
Error	$3(n-1)$

En este ejemplo

$$\bar{X}_1 = \overline{\text{hembras}}$$

$$\bar{X}_2 = \overline{\text{machos}}$$

$$\bar{X}_3 = \overline{\text{mezcla}}$$

El contraste entre machos vs hembras es

$$D_{MH} = \overline{\text{hembras}} - \overline{\text{machos}} = \bar{X}_1 - \bar{X}_2.$$

La media de la mezcla es

$$\frac{\overline{\text{hembras}} + \overline{\text{machos}}}{2} = \frac{\bar{X}_1 + \bar{X}_2}{2}$$

Por lo que el contraste de crianza juntas vs separadas es

$$D_{JS} = \frac{\overline{\text{hembras}} + \overline{\text{machos}}}{2} - \overline{\text{mezcla}} = \frac{\bar{X}_1 + \bar{X}_2}{2} - \bar{X}_3.$$

Aunque, en estos dos ejemplos, se ha descompuesto la suma de cuadrados de tratamientos ( $SC$ ) por coeficientes ortogonales, en el primer ejemplo podría haberse descompuesto directamente introduciendo los dos factores  $C$  y  $P$  en vez de el único factor tratamientos. En cambio en el segundo ejemplo queda más patente la ventaja de esta descomposición puesto que no se trata de aumentar el modelo con otro factor más, sino de profundizar en el problema bajo un punto de vista que no es inmediato.

Como se ve, estas pruebas de separación ortogonales de medias pueden formular y responder a tantas preguntas independientes como grados de libertad se tenga para los tratamientos. Aunque hay que hacer notar que las preguntas deben plantearse antes de que se realicen los experimentos.

Para realizar la  $SC$  de un contraste de este tipo, con un solo grado de libertad, puede calcularse a partir del conjunto de coeficientes cuya suma es cero de la siguiente manera: el sumatorio de cada tratamiento ( $X_i$ ) se multiplica por el coeficiente correspondiente ( $c_i$ ) que se le ha asignado en el contraste, y se suman todos los niveles elevando al cuadrado esa suma y esto se divide por la suma de cuadrados de los coeficientes multiplicada por el número de términos que hay dentro de cada nivel. La fórmula es

$$SC_{\text{contraste}} = \frac{(\sum_i c_i X_i)^2}{n \sum_i c_i^2}$$

cada contraste se analiza con un  $gl$ .

### Ejemplo.-

Se quiere saber si hay diferencia en la producción de piezas hechas por cuatro trabajadores diferentes. Los datos son

	$T_1$	$T_2$	$T_3$	$T_4$
	97	100	80	50
	80	50	85	55
	100	60	80	60
	103	100	70	60
	70	90	85	75
$X_i$	450	400	400	300
$\bar{X}_i$	90	80	80	60

El análisis es

$FV$	$gl$	$SC$	$CM$	$F_o$	$ECM$
<i>Tratamientos</i>	3	2375	791.667	3.61*	$\sigma^2 + 5\sigma_T^2$
<i>Error</i>	16	3518	219.875		$\sigma^2$
<i>total</i>	19	5893			

(este análisis se realiza tal como se ha visto en el epígrafe *Prueba de Hipótesis en el modelo de una vía*).

Al ser el factor trabajador significativo al  $\alpha=0.05$ , indica que no producen igual todos y que hay unos que rinden más que otros. Pero antes de realizar el análisis se sabía que dos son de un sexo ( $H$ ) y dos son del otro ( $M$ ) y unos tienen coeficiente de inteligencia alto ( $A$ ) y otros bajo ( $B$ ). Los sexo y coeficientes de inteligencia se distribuye de la siguiente forma

	$Sexo$	$CI$
$T_1$	$H$	$A$
$T_2$	$M$	$A$
$T_3$	$H$	$B$
$T_4$	$M$	$B$

Los contrastes más lógicos serán

- 1) Influencia del sexo en la producción.
- 2) Influencia del coeficiente de inteligencia.
- 3) Si existe interacción entre sexos y coeficiente de inteligencia.

Los coeficientes ( $c$ ) para estos contrastes son

	$c_1$	$c_2$	$c_3$	$c_4$	$\sum c_j^2$
<i>Sexo</i>	1	-1	1	-1	4
<i>CI</i>	1	1	-1	-1	4
<i>Sexo CI</i>	1	-1	-1	1	4

Las sumas de cuadrados son

$$SC_{\text{sexo}} = \frac{(450 - 400 + 400 - 300)^2}{n \sum c^2} = 1125$$

siendo  $n \sum c^2 = 5 \times 4 = 20$

$$SC_{CI} = \frac{(450 + 400 - 400 - 300)^2}{n \sum c^2} = 1125$$

$$SC_{\text{interacción}} = \frac{(450 - 400 - 400 + 300)^2}{n \sum c^2} = 125$$

El análisis sería, ahora

<i>FV</i>	<i>gl</i>	<i>SC</i>	<i>CM</i>	<i>ECM</i>	<i>F</i>
<i>Trabajadores</i>	3	2375	791.667	$\sigma^2 + t \sigma_T^2$	3.61*
<i>Sexo</i>	1	1125	1125	$\sigma^2 + c_1 \sigma_{Sx}^2$	5.12*
<i>CI</i>	1	1125	1125	$\sigma^2 + c_2 \sigma_{CI}^2$	5.12*
<i>Sex CI</i>	1	125	125	$\sigma^2 + c_3 \sigma_{SC}^2$	0.568ns
<i>Error</i>	16	3518	219.875	$\sigma^2$	
<i>total</i>	19	5893			

De este nuevo análisis se concluye que los sexos de los trabajadores y su coeficiente de inteligencia influyen en la producción de piezas, no habiendo por el contrario interacción entre estos factores.

En este problema, al igual que se han hecho estos contrastes, podrían haberse realizado otros, por ejemplo, comparar  $M$  de  $CI$  alto ( $MA$ ) con  $H$  de  $CI$  bajo ( $CB$ ) y viceversa,  $HA$  con  $MB$ . Los coeficientes de estos contrastes serán



	$C_1$	$C_2$	$C_3$	$C_4$
$C_1$ ) MA vs HB	0	1	-1	0
$C_2$ ) HA vs MB	1	0	0	-1

$$SC_{C_1} = \frac{(400 - 400)^2}{2 \times 5} = 0$$

$$SC_{C_2} = \frac{(450 - 300)^2}{2 \times 5} = 2250$$

El análisis será

FV	gl	SC	CM	ECM	F
Trabajadores	3	2375	791.667		3.61*
$C_1$	1	0	0	$\sigma^2 + C_1 \sigma_{C_1}^2$	1.00ns
$C_2$	1	2250	2250	$\sigma^2 + C_2 \sigma_{C_2}^2$	10.23**
Resto	1	125	125	$\sigma^2 + C_3 \sigma_{R\theta}^2$	0.57ns
Error	16	3518	219.875		
total	19	5893			

De estos resultados se concluye que entre  $M$  de alto  $CI$  y  $H$  de bajo  $CI$  no hay ninguna diferencia de producción, debiéndose la mayor parte de la variabilidad encontrada al segundo contraste. El resto (parte que ha quedado sin descomponer:  $SC_{trabaja-C_1-C_2}$ ) tampoco presenta significación.

#### Archivo del programa SAS (C10-1.SAS).-

```

title 'Pruebas planeadas de trabajadores y producción piezas';
options ls=75 ps=60;
data orto;
infile 'c10-1.dat';
input trabaj $ prod @@;
proc glm;
class trabaj;
model prod = trabaj / ss3;
contrast 'sexo'      trabaj  1 -1  1 -1;
contrast 'Coef Inte' trabaj  1  1 -1 -1;
contrast 'sexo x CI' trabaj  1 -1 -1  1;
contrast 'C1) MA vs HB' trabaj  0  1 -1  0;
contrast 'C2) HA vs MB' trabaj  1  0  0 -1;
contrast 'resto'     trabaj  1 -1 -1  1;
run;

```

### Archivo de datos (C10-1.DAT)-

1	97	2	100	3	80	4	50
1	80	2	50	3	85	4	55
1	100	2	60	3	80	4	60
1	103	2	100	3	70	4	60
1	70	2	90	3	85	4	75

### Archivo de resultados (C10-1.LST)-

Dependent Variable: PROD					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2375.00000	791.66667	3.60	0.0368
Error	16	3518.00000	219.87500		
Corrected Total	19	5893.00000			
	R-Square	C.V.	Root MSE		PROD Mean
	0.403021	19.13314	14.8282		77.5000
Source	DF	Type III SS	Mean Square	F Value	Pr > F
TRABAJ	3	2375.00000	791.66667	3.60	0.0368
Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
sexo	1	1125.00000	1125.00000	5.12	0.0380
Coef Inte	1	1125.00000	1125.00000	5.12	0.0380
sexo x CI	1	125.00000	125.00000	0.57	0.4618
C1) MA vs HB	1	0.00000	0.00000	0.00	1.0000
C2) HA vs MB	1	2250.00000	2250.00000	10.23	0.0056
resto	1	125.00000	125.00000	0.57	0.4618

Se concluye que los sexos de los trabajadores y su coeficiente de inteligencia influyen en la producción de piezas, no habiendo por el contrario interacción entre estos factores. Y, así mismo, se concluye que entre *M* de alto *CI* y *H* de bajo *CI* no hay ninguna diferencia de producción, debiéndose la mayor parte de la variabilidad encontrada al segundo contraste. El resto (parte que ha quedado sin descomponer:  $SC_{trabaja-C_1-C_2}$ ) tampoco presenta significación.

### Modelos con más de un factor.-

Si se tiene un modelo con más de un factor y se quiere hacer pruebas planeadas de uno o unos factores, todo se haría lo mismo teniendo en cuenta, solamente, que habría que utilizar como término de error el mismo del factor que se esta descomponiendo.

### Ejemplo.-

Para estudiar los niveles de cobre existente en diferentes tipos de lesiones, se tomó una muestra aleatoria de ocho ratas a las que se sometió a quemaduras en una zona de la piel. Tomadas muestras de las zonas quemadas y no quemadas (control) de cada una de las ocho ratas, se midió en ellas el nivel de cobre, teniendo en cuenta,

también, en cada zona quemada, tres subzonas según su distancia al foco de la quemadura. Los datos son

Rata	Control	Subzona 1	Subzona 2	Subzona 3
1	15.031	17.910	16.531	16.660
2	23.531	26.501	35.423	34.111
3	20.100	45.151	22.630	20.176
4	26.142	30.050	33.442	30.567
5	26.546	35.193	37.608	30.119
6	19.403	35.109	30.432	24.631
7	16.431	31.841	23.197	20.115
8	16.003	21.369	20.756	18.441

- ¿Existe diferencia entre las cuatro zonas?
- ¿Existe diferencia entre las zonas quemadas y no quemadas?
- ¿Existe diferencia entre las subzonas quemadas?

La respuesta a la primera pregunta la da la prueba  $F$  del factor *zona* del análisis de dos vías. Hay que utilizar el modelo de dos vías para que el término de error sea lo más pequeños posible y, por lo tanto, el análisis sea más preciso.

La respuesta a la segunda pregunta la da la prueba planeada de coeficientes

*Zona 3 -1 -1 -1*

Y la respuesta a la tercera pregunta viene dada por diferencia, ya que si a la suma de cuadrados debida a *Zonas* se le resta la suma de cuadrados del contraste de la expresión anterior, queda la suma de cuadrados debida a las subzonas quemadas, con dos grados de libertad. Si esta fuente de variación fuera significativa y se quisiera saber entre cuáles zonas existen diferencias, habría que hacer un *contraste múltiple*, que se verán más adelante.

#### Archivo del programa SAS (C10-2.SAS).-

```

title 'Contraste ortogonal de subzonas quemadas y no quemadas';
options ls=75 ps=60;
data orto;
infile 'c10-2.dat';
input rata $ zona $ cobre @@;
proc glm;
  class rata zona;
  model cobre = rata zona;
  contrast 'quemad vs no quemada' zona 3 -1 -1 -1;
run;

```

#### Archivo de datos (C10-2.DAT).-

R1 C 15.031	R1 Z1 17.910	R1 Z2 16.531	R1 Z3 16.660
R2 C 23.531	R2 Z1 26.501	R2 Z2 35.423	R2 Z3 34.111
R3 C 20.100	R3 Z1 45.151	R3 Z2 22.630	R3 Z3 20.176
R4 C 26.142	R4 Z1 30.050	R4 Z2 33.442	R4 Z3 30.567

R5 C 26.546	R5 Z1 35.193	R5 Z2 37.608	R5 Z3 30.119
R6 C 19.403	R6 Z1 35.109	R6 Z2 30.432	R6 Z3 24.631
R7 C 16.431	R7 Z1 31.841	R7 Z2 23.197	R7 Z3 20.115
R8 C 16.003	R8 Z1 21.369	R8 Z2 20.756	R8 Z3 18.441

**Archivo de resultados (C10-2.LST).-**

Dependent Variable: COBRE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	1323.00899	132.30090	5.61	0.0004
Error	21	495.14474	23.57832		
Corrected Total	31	1818.15373			
	R-Square	C.V.	Root MSE	COBRE Mean	
	0.727666	18.92274	4.85575	25.6609	
Source	DF	Type III SS	Mean Square	F Value	Pr > F
RATA	7	881.679179	125.954168	5.34	0.0013
ZONA	3	441.329810	147.109937	6.24	0.0034
Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
quemad vs no quemada	1	295.408683	295.408683	12.53	0.0019

Tal como se ha dicho en el planteamiento del problema, para saber si existe diferencia entre las subzonas quemadas, se hace por diferencia de la siguiente manera

<i>FV</i>	<i>gl</i>	<i>SC</i>	<i>CM</i>	<i>F</i>
<i>Rata</i>	7	881.6792	125.9542	5.34**
<i>Zonas</i>	3	441.3298	147.1099	6.24**
<i>Quem vs no Quem entre quemadas</i>	1	295.4087	295.4087	12.53**
	2	145.9211	72.9605	3.094ns
<i>Error</i>	21	495.1447	23.5783	
<i>total</i>	31	1818.1537		

Como no existe diferencia entre las subzonas quemadas, no hay que hacer el contraste múltiple, puesto que no existe evidencia de que las tres zonas tengan diferente cantidad de cobre.

Si se tiene un modelo con más de un factor y el término de error que se está utilizando para el factor que se va a descomponer es diferente al *Error* (por ejemplo, una interacción), hay que utilizar el mismo término de error para las pruebas planeadas.

**Ejemplo.-**

Se han probado cuatro tratamientos en cuatro rebaños elegidos al azar. Los cuatro tratamientos son cuatro piensos que tienen las siguientes características

	Proteínas	Energía
Pienso 1	20%	3000
Pienso 2	20%	3500
Pienso 3	22%	3000
Pienso 4	22%	3500

Los datos son

	Tratamientos			
	T <sub>1</sub>	T <sub>2</sub>	T <sub>3</sub>	T <sub>4</sub>
R <sub>1</sub>	73	82	45	55
	33	72	85	85
R <sub>2</sub>	12	53	32	31
	22	83	22	51
R <sub>3</sub>	69	64	34	64
	19	94	34	34
R <sub>4</sub>	27	66	18	82
	87	76	48	72

### Archivo del programa SAS (C10-3.SAS).-

```

title 'Comparación planeada de tratamientos, modelo mixto';
options ls=75 ps=60;
data medias;
infile 'c10-3.dat';
input rebano $ trata $ prod @@;
proc glm;
class rebano trata ;
model prod = rebano trata rebano*trata /ss3;
contrast 'Energia' trata -1 1 -1 1 / E=rebano*trata;
contrast 'Proteina' trata -1 -1 1 1 / E=rebano*trata;
contrast 'Ener/Prot' trata 1 -1 -1 1 / E=rebano*trata;
test H=trata E=rebano*trata;
run;

```

### Archivo de datos (C10-3.DAT).-

```

R1 T1 73 R1 T2 82 R1 T3 45 R1 T4 55
R2 T1 12 R2 T2 53 R2 T3 32 R2 T4 31
R3 T1 69 R3 T2 64 R3 T3 34 R3 T4 64
R4 T1 27 R4 T2 66 R4 T3 18 R4 T4 82
R1 T1 33 R1 T2 72 R1 T3 85 R1 T4 85
R2 T1 22 R2 T2 83 R2 T3 22 R2 T4 51
R3 T1 19 R3 T2 94 R3 T3 34 R3 T4 34
R4 T1 87 R4 T2 76 R4 T3 48 R4 T4 72

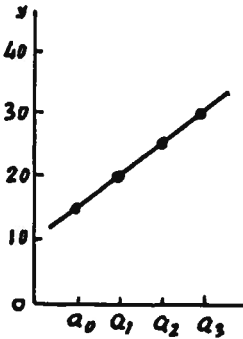
```

Archivo de resultados (C10-3.LST)-

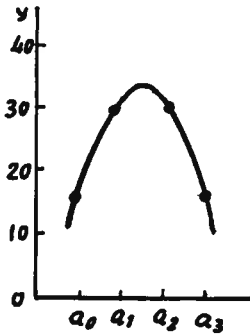
Dependent Variable: PROD					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	11575.5000	771.7000	1.68	0.1570
Error	16	7350.0000	459.3750		
Corrected Total	31	18925.5000			
	R-Square	C.V.	Root MSE	PROD Mean	
	0.611635	39.78290	21.4330	53.8750	
Source	DF	Type III SS	Mean Square	F Value	Pr > F
REBANO	3	3476.50000	1158.83333	2.52	0.0945
TRATA	3	5977.50000	1992.50000	4.34	0.0204
REBANO*TRATA	9	2121.50000	235.72222	0.51	0.8443
Tests of Hypotheses using the Type III MS for REBANO*TRATA as an error term					
Source	DF	Type III SS	Mean Square	F Value	Pr > F
TRATA	3	5977.50000	1992.50000	8.45	0.0055
Tests of Hypotheses using the Type III MS for REBANO*TRATA as an error term					
Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Energia	1	5100.50000	5100.50000	21.64	0.0012
Tests of Hypotheses using the Type III MS for REBANO*TRATA as an error term					
Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Proteina	1	612.50000	612.50000	2.60	0.1414
Tests of Hypotheses using the Type III MS for REBANO*TRATA as an error term					
Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Ener/Prot	1	264.50000	264.50000	1.12	0.3171

## Contrastes ortogonales con factores cuantitativos.-

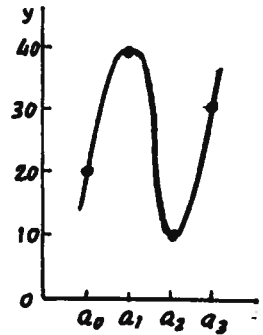
Un tipo de comparaciones planeadas muy interesantes son los contrastes ortogonales de los tratamientos cuando éstos tienen niveles numéricos igualmente espaciados. En este caso, lo lógico es estudiar la tendencia que presenta la variable analizada al aumentar progresivamente los niveles del tratamiento. Esta tendencia puede ser *lineal*, *cuadrática*, *cúbica*, etc., (ver figura siguiente) y una vez establecida de qué tipo es nos ayudará a interpretar los resultados.



todos los efectos lineales. Los otros componentes nulos



todos los efectos cuadráticos. Los otros componentes nulos



todos los efectos cúbicos. Los otros componentes nulos

El cálculo de los coeficientes para estos contrastes se puede realizar de una forma lógica, teniendo en cuenta el tipo de línea a la que se quiere ajustar los puntos obtenidos. En el caso de rectas y parábolas es relativamente fácil calcularlas, pero en curvas de orden superior es más complejo su estudio si bien hay tablas que ayudan en dicho cometido. Los coeficientes para las sumas cuadrados se encuentran en la Tabla 6 y no hay más que aplicarlos directamente para la obtención de las SC.

Una vez contrastado que los efectos de los niveles del factor son significativos, se puede determinar la significación de los efectos lineal, cuadrático, cúbico, etc., (uno menos que el número de niveles). Para ello se necesita conocer qué parte de la variación debida al factor procede del efecto lineal, qué parte procede del efecto cuadrático, etc.

Para mayor claridad expliquémoslo con un ejemplo.

**Ejemplo.-**

Sean, por ejemplo, cuatro dosis: 15, 20, 25 y 30, habiéndose obtenido las siguientes respuestas

	15	20	22	30
0	0	1	2	3
0	0	2	3	4
1	1	1	2	3
1	1	1	2	3
$\Sigma$	2	5	9	13

$$\Sigma T^2 = \frac{1}{n} \Sigma_i X_i^2 = \frac{1}{4} (2^2 + 5^2 + 9^2 + 13^2) = 69.75$$

$$TC = \frac{X_{..}^2}{N} = \frac{29^2}{16} = 52.56$$

$$\Sigma X^2 = \Sigma_{ij} X_{ij}^2 = 0^2 + 0^2 + \dots + 3^2 = 73$$

FV	gl	SC	CM	F <sub>o</sub>	ECM
Tratamientos	3	17.19	5.73	21.16**	$\sigma^2 + 4 \sigma_T^2$
Error	12	3.25	0.27		$\sigma^2$
total	15	20.44			

La F de contraste es  $F_{(3,12; 0.05)} = 4.49$

Al ser la F<sub>o</sub> significativa, indica que los efectos de los tratamientos son efectivos, es decir, existe un efecto dosis. Para saber que tipo de efecto tiene la dosis, se va a estimar la lineal, la cuadrática y la cúbica

Los coeficientes dados en la Tabla 6 para dichos efectos y cuatro puntos son

	c <sub>1</sub>	C <sub>2</sub>	c <sub>3</sub>	c <sub>4</sub>	$\Sigma c_i^2$
Lineal C <sub>1</sub>	-3	-1	1	3	20
Cuadrático C <sub>2</sub>	+1	-1	-1	1	4
Cúbico C <sub>3</sub>	-1	3	-3	1	20

Las SC de las diferentes respuestas son

$$SC_{C_i} = \frac{(c_1 X_1 + c_2 X_2 + c_3 X_3 + c_4 X_4)^2}{t \Sigma c_i^2}$$

Por lo que se tiene que la sumas de cuadrados de las tres respuestas son

$$SC_{Lineal} = \frac{[(-3 \times 2) + (-1 \times 5) + (1 \times 9) + (3 \times 13)]^2}{4 \times 20} = 17.11$$



$$SC_{\text{Cuadrática}} = \frac{[(1 \times 2) + (-1 \times 5) + (-1 \times 9) + (1 \times 13)]^2}{4 \times 4} = 0.063$$

$$SC_{\text{Cúbica}} = \frac{[(-1 \times 2) + (3 \times 5) + (-3 \times 9) + (1 \times 13)]^2}{4 \times 20} = 0.0125$$

Quedando el análisis así

FV	gl	SC	CM	ECM	F
Tratamiento	3	17.1875	5.7292	$\sigma^2 + t \sigma_T^2$	21.16**
Lineal	1	17.1125	17.1125	$\sigma^2 + C_1 \sigma_{\text{Lineal}}^2$	63.37**
Cuadrática	1	0.0625	0.0625	$\sigma^2 + C_2 \sigma_{\text{Cuadrá}}^2$	0.24ns
Cúbica	1	0.0125	0.0125	$\sigma^2 + C_3 \sigma_{\text{Cúbica}}^2$	0.05ns
Error	12	20.4375	0.2708	$\sigma^2$	
total	15	20.4375			

Estos resultados indican que las respuestas aumentan linealmente al aumentar las dosis de abono. No hay respuesta cuadrática ni cúbica.

#### Archivo del programa SAS (C10-4.SAS).-

```

title 'Contrastes ortogonales (lineal, cuadrático, cúbico) de
respuestas a dosis';
options ls=75 ps=60;
data orto;
infile 'c10-4.dat';
input dosis resp @@;
proc glm;
class dosis;
model resp = dosis;
contrast 'lineal'      dosis -3 -1 1 3;
contrast 'Cuadratica'  dosis 1 -1 -1 1;
contrast 'cubica'     dosis -1 3 -3 1;
run;

```

#### Archivo de datos (C10-4.DAT).-

```

15 0   20 1   25 2   30 3
15 0   20 2   25 3   30 4
15 1   20 1   25 2   30 3
15 1   20 1   25 2   30 3

```

#### Archivo de resultados (C10-4.LST).-

Dependent Variable: RESP						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	3	17.1875000	5.7291667	21.15	0.0001	
Error	12	3.2500000	0.2708333			
Corrected Total	15	20.4375000				

	R-Square	C.V.	Root MSE	RESP Mean	
	0.840979	28.71263	0.52042	1.81250	
Source	DF	Type I SS	Mean Square	F Value	Pr > F
DOSIS	3	17.1875000	5.7291667	21.15	0.0001
Source	DF	Type III SS	Mean Square	F Value	Pr > F
DOSIS	3	17.1875000	5.7291667	21.15	0.0001
Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
lineal	1	17.1125000	17.1125000	63.18	0.0001
Cuadrática	1	0.0625000	0.0625000	0.23	0.6396
cubica	1	0.0125000	0.0125000	0.05	0.8335

**Contrastes ortogonales con modelos factoriales de dos factores, ambos cuantitativos.-**

Supóngase el factor  $T$  con seis niveles y el factor  $R$  con cuatro niveles en un diseño equilibrado con  $n$  observaciones por celda. Siendo el modelo

$$X_{ijk} = \mu + T_i + R_j + TR_{ij} + e_{ijk}$$

Los coeficientes dados por las Tabla 6 para los seis puntos del factor  $T$  son

	$C_1$	$C_2$	$C_3$	$C_4$	$C_5$	$C_6$	$\sum C_i^2$
Lineal $C_1$	-5	-3	-1	1	3	5	70
Cuadrático $C_2$	+5	-1	-4	-4	-1	5	84
Cúbico $C_3$	-5	7	4	-4	-7	5	180

Las  $SC$  de las diferentes respuestas del factor  $T$  se realizarían igual que antes, esto es

$$SC_{C_i} = \frac{(c_1 X_{1..} + c_2 X_{2..} + c_3 X_{3..} + c_4 X_{4..} + c_5 X_{5..} + c_6 X_{6..})^2}{r n \sum c_i^2}$$

Al tener seis niveles se hubiera podido calcular la suma de cuadrados de la tendencia cuártica y de la tendencia quinta, pero como la interpretación de estas tendencias puede ser complicada, mejor se calcula el residuo de la siguiente manera

$$R = SC_T - (SC_{Lineal} + SC_{Cuadrática} + SC_{Cúbica})$$

Si fuera significativo este residuo es porque la respuesta es de cuarto o quinto orden.

Para el segundo factor ( $R$ ), los coeficientes dados por la Tabla 6 para los cuatro niveles son

	$c_1$	$c_2$	$c_3$	$c_4$	$\sum c_i^2$
Lineal $C_1$	-3	-1	1	3	20
Cuadrático $C_2$	+1	-1	-1	1	4
Cúbico $C_3$	-1	3	-3	1	20

Las SC de las diferentes respuestas son

$$SC_{C_i} = \frac{(c_1 X_{.1} + c_2 X_{.2} + c_3 X_{.3} + c_4 X_{.4})^2}{t n \sum c_i^2}$$

Al ser cuatro el número de niveles del factor  $R$  no queda residuo.

### Descomposición de la suma de cuadrados de la interacción.-

Se ha descompuesto la suma de cuadrados de cada efecto principal en tantas componentes como grados de libertad presenta. De la misma manera se puede descomponer la interacción en tantos componentes como grados de libertad tenga ésta. En el caso del modelo que se está estudiando serían 15 componentes. No todas van a ser interesantes, por lo que nos podemos limitar a las más interesantes, como pueden ser

*Respuesta Lineal de T × Respuesta Lineal de R*  
*Respuesta Lineal de T × Respuesta Cuadrática de R*  
*Respuesta Cuadrática de T × Respuesta Lineal de R*

Para obtener las sumas de cuadrados de la componente de la interacción *lineal x lineal*, se hallan los coeficientes de la siguiente forma: se elabora una tabla de doble entrada de dimensión  $txr$ , los valores de entrada de las columnas son los coeficientes ortogonales de la componente lineal del factor  $R$  y los valores de entrada de las filas son los coeficientes ortogonales de la componente lineal del factor  $T$ ; en cada casilla se coloca el producto de los coeficientes ortogonales de la fila y columna donde esta colocada la casilla en cuestión, así como las sumas de las observaciones de dicha casilla. La tabla quedaría así

		Factor R			
		-3	-1	1	3
Factor T	-5	15 $X_{11.}$	5 $X_{12.}$	-5 $X_{13.}$	-15 $X_{14.}$
	-3	9 $X_{21.}$	3 $X_{22.}$	-3 $X_{23.}$	-9 $X_{24.}$
	-1	3 $X_{31.}$	1 $X_{32.}$	-1 $X_{33.}$	-3 $X_{34.}$
	1	-3 $X_{41.}$	-1 $X_{42.}$	1 $X_{43.}$	3 $X_{44.}$
	3	-9 $X_{51.}$	-3 $X_{52.}$	3 $X_{53.}$	9 $X_{54.}$
	5	-15 $X_{61.}$	-5 $X_{62.}$	5 $X_{63.}$	15 $X_{64.}$

$$\sum c_i^2 = 1364$$

La SC de la respuesta *lineal x lineal* es, por tanto, el sumatorio de los productos del coeficiente y suma de cada casilla de la tabla anterior, elevado al cuadrado y dividido por el número total de datos por el sumatorio de los cuadrados de los coeficientes, esto es

$$SC_{\text{Lineal x lineal}} = \frac{(15 X_{11.} + 5 X_{12.} - 5 X_{13.} \dots - 5 X_{62.} + 5 X_{63.} + 15 X_{64.})^2}{tm \sum c_i^2} \quad \text{Para}$$

la interacción *lineal x cuadrática*, se actúa de la misma manera. Los coeficientes son

		Factor R			
		1	-1	-1	1
Factor T	-5	-5 $X_{11.}$	5 $X_{12.}$	5 $X_{13.}$	-5 $X_{14.}$
	-3	-3 $X_{21.}$	3 $X_{22.}$	3 $X_{23.}$	-3 $X_{24.}$
	-1	-1 $X_{31.}$	1 $X_{32.}$	1 $X_{33.}$	-1 $X_{34.}$
	1	1 $X_{41.}$	-1 $X_{42.}$	-1 $X_{43.}$	1 $X_{44.}$
	3	3 $X_{51.}$	-3 $X_{52.}$	-3 $X_{53.}$	3 $X_{54.}$
	5	5 $X_{61.}$	-5 $X_{62.}$	-5 $X_{63.}$	5 $X_{64.}$

$$\sum c_i^2 = 280$$

La SC de la respuesta *lineal x cuadrática* es

$$SC_{\text{Lineal} \times \text{Cuadrática}} = \frac{(-5 X_{11} + 5 X_{12} + 5 X_{13} \dots - 5 X_{62} - 5 X_{63} + 5 X_{64})^2}{trn \sum c_i^2} \quad \text{Para}$$

la interacción *cuadrática x lineal*, Los coeficientes son

		Factor R			
		-3	-1	1	3
Factor T	5	-15 $X_{11}$	-5 $X_{12}$	5 $X_{13}$	15 $X_{14}$
	-1	3 $X_{21}$	1 $X_{22}$	-1 $X_{23}$	-3 $X_{24}$
	-4	12 $X_{31}$	4 $X_{32}$	-4 $X_{33}$	-12 $X_{34}$
	-4	12 $X_{41}$	4 $X_{42}$	-4 $X_{43}$	-12 $X_{44}$
	-1	3 $X_{51}$	1 $X_{52}$	-1 $X_{53}$	-3 $X_{54}$
	5	-15 $X_{61}$	-5 $X_{62}$	5 $X_{63}$	15 $X_{64}$

$$\sum c_i^2 = 680$$

La SC de la respuesta *cuadrática x lineal* es:

$$SC_{\text{Cuadrática} \times \text{lineal}} = \frac{(-15 X_{11} - 5 X_{12} + 5 X_{13} \dots - 5 X_{62} + 5 X_{63} + 15 X_{64})^2}{trn \sum c_i^2}$$

El resto de las interacciones serán

$$R = SC_{\text{trata} \times \text{raza}} - (SC_{\text{lineal} \times \text{lineal}} + SC_{\text{lineal} \times \text{cuadrática}} + SC_{\text{cuadrática} \times \text{lineal}})$$

### Ejemplo.-

Se desea estudiar el tipo de influencia de dos factores en el gasto familiar de un producto, analizando si la influencia de los efectos principales es lineal, cuadrática o cúbica, como asimismo distintas componentes de la interacción total. Los factores son los ingresos de las familias y la densidad de población de las ciudades en que viven. Se realizó dos repeticiones y la variable medida son miles de pesetas gastadas en dicho producto. Los datos son

		<i>Ingresos</i>			
		100	200	300	400
<i>Densidad</i>	100	5.30	5.41	5.43	5.59
		5.27	5.33	5.43	5.56
	200	5.40	5.50	5.64	5.70
		5.35	5.53	5.63	5.76
	300	5.49	5.67	5.76	5.89
		5.53	5.65	5.78	5.86
	400	5.59	5.79	5.90	5.98
		5.57	5.82	5.93	6.05
	500	5.14	5.25	5.31	5.49
		5.13	5.20	5.32	5.47
	600	5.04	5.12	5.26	5.34
		5.10	5.12	5.24	5.37

los sumatorios son

		<i>Ingresos</i>				
		100	200	300	400	$X_{i.}$
<i>Densidad</i>	100	10.57	10.74	10.86	11.15	43.32
	200	10.75	11.03	11.27	11.46	44.51
	300	11.02	11.32	11.54	11.45	45.63
	400	11.16	11.61	11.83	12.03	46.63
	500	10.27	10.45	10.63	10.96	42.31
	600	10.14	10.35	10.50	10.71	41.65
$X_{.j}$		63.91	65.45	66.63	68.06	264.05

Las componentes del factor *ingresos* son

$$SC_{\text{Lineal}} = \frac{(-5 \times 43.32 - 3 \times 44.51 - 1 \times 45.63 + 1 \times 46.63 + 3 \times 42.31 + 5 \times 41.65)^2}{4 \times 2 \times 70} =$$

$$= 0.3475$$

$$SC_{\text{Cuadrática}} = \frac{(5 \times 43.32 - 1 \times 44.51 - 4 \times 45.63 - 4 \times 46.63 - 1 \times 42.31) + 5 \times 41.65)^2}{4 \times 2 \times 84} =$$

$$= 1.4310$$

$$SC_{\text{Cúbica}} = \frac{(-5 \times 43.32 + 7 \times 44.51 + 4 \times 45.63 - 4 \times 46.63 - 7 \times 42.31 + 5 \times 41.65)^2}{4 \times 2 \times 180} =$$

$$= 0.3885$$

$$\text{Resto} = 2.1735 - (0.3475 + 1.4310 + 0.0065) = 0.3885$$

Las componentes del factor *densidad* son

$$SC_{\text{Lineal}} = \frac{(-3 \times 63.91 - 1 \times 65.45 + 1 \times 66.3 + 3 \times 68.06)^2}{6 \times 2 \times 20} = 0.7741$$

$$SC_{\text{Cuadrática}} = \frac{(1 \times 63.91 - 1 \times 65.45 - 1 \times 66.63 + 1 \times 68.06)^2}{6 \times 2 \times 4} = 0.00025$$

$$SC_{\text{Cúbica}} = \frac{(-1 \times 63.91 + 3 \times 65.45 - 3 \times 66.63 + 1 \times 68.06)^2}{6 \times 2 \times 20} = 0.00155$$

Al tener este factor cuatro niveles y al haber llegado hasta la componente cúbica, no queda resto.

Las componentes de la interacción que se van a estudiar son

$$SC_{\text{Lineal} \times \text{Lineal}} = \frac{(15 \times 10.57 + 5 \times 10.74 + \dots + 5 \times 10.50 + 15 \times 10.71)^2}{2 \times 1364} = 0.000035$$

$$SC_{\text{Lineal} \times \text{Cuadrática}} = \frac{(-5 \times 10.57 + 5 \times 10.74 + \dots - 5 \times 10.50 + 5 \times 10.71)^2}{2 \times 280} = 0.000079$$

$$SC_{\text{Cuadrática} \times \text{Lineal}} = \frac{(-15 \times 10.57 - 5 \times 10.74 + \dots + 5 \times 10.50 + 15 \times 10.71)^2}{2 \times 1680} = 0.013481$$

$$\text{Resto} = 0.0320 - (0.000035 + 0.000079 + 0.013481) = 0.018405$$

El análisis es

<i>FV</i>	<i>gl</i>	<i>SC</i>	<i>CM</i>	<i>F</i>
<i>Ingresos</i>	5	2.3343	0.4669	598.59***
<i>Lineal</i>	1	0.3475	0.3475	445.51***
<i>Cuadrática</i>	1	1.4310	1.4320	1834.61***
<i>Cúbica</i>	1	0.0065	0.0065	8.33**
<i>Resto</i>	2	0.5493	0.2747	352.18***
<i>Densidad</i>	3	0.7759	0.2586	331.54***
<i>Lineal</i>	1	0.7741	0.7741	992.39***
<i>Cuadrática</i>	1	0.00025	0.00025	0.32ns
<i>Cúbica</i>	1	0.00155	0.00155	1.99ns
<i>Ingr x Dens</i>	15	0.1929	0.0129	16.54*
<i>Lineal</i>	1	0.000035	0.000035	0.04ns
<i>Cuadrática</i>	1	0.000079	0.000079	0.10ns
<i>Cúbica</i>	1	0.013481	0.013481	17.28***
<i>Resto</i>	12	0.018405	0.001534	1.96ns
<i>Error</i>	24	0.0186	0.00078	
<i>total</i>	47	3.1604		

**Archivo del programa SAS (C10-5.SAS)-**

```

title 'Contrastes ortogonales (line., cuadr., cubi.) de dos
factores';
options ls=75 ps=60;
data orto;
infile 'c10-5.dat';
do ingresos=1 to 6;
  do densidad=1 to 4;
    do repe=1 to 2;
      input gasto @@;output;
    end;
  end;
end;

proc glm;
class ingresos densidad;
model gasto = ingresos densidad ingresos*densidad;
contrast 'Ingresos, lineal' ingresos -5 -3 -1 1 3 5;
contrast 'Ingresos, cuadratica' ingresos 5 -1 -4 -4 -1 5;
contrast 'Ingresos, cubica' ingresos -5 7 4 -4 -7 5;
contrast 'Densidad, lineal' densidad -3 -1 1 3;
contrast 'Densidad, cuadratica' densidad 1 -1 -1 1;
contrast 'Densidad, cubica' densidad -1 3 -3 1;
contrast 'lineal*lineal' ingresos*densidad 15 5 -5 -15 9 3
-3 -9 3 1 -1 -3 -3 -1 1 3 -9 -3 3 9 -15 -5 5 15;
contrast 'lineal*cuadratica' ingresos*densidad -5 5 5 -5 -3
3 3 -3 -1 1 1 -1 1 -1 -1 1 3 -3 -3 3 5 -5 -5 5;
contrast 'cuadratic*lineal' ingresos*densidad -15 -5 5 15 3
1 -1 -3 12 4 -4 -12 12 4 -4 -12 3 1 -1 -3 -15 -5 5 15;
run;

```

**Archivo de datos (C10-5.DAT)-**

5.30	5.27	5.41	5.33	5.43	5.43	5.59	5.56
5.40	5.35	5.50	5.53	5.64	5.63	5.70	5.76
5.49	5.53	5.67	5.65	5.76	5.78	5.89	5.86
5.59	5.57	5.79	5.82	5.90	5.93	5.98	6.05
5.14	5.13	5.25	5.20	5.31	5.32	5.49	5.47
5.04	5.10	5.12	5.18	5.26	5.24	5.34	5.37

**Archivo de resultados (C10-5.LST)-**

Dependent Variable: GASTO						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	23	3.14219792	0.13661730	175.81	0.0001	
Error	24	0.01865000	0.00077708			
Corrected Total	47	3.16084792				
	R-Square	C.V.	Root MSE	GASTO Mean		
	0.994100	0.506744	0.02788	5.50104		
Source	DF	Type III SS	Mean Square	F Value	Pr > F	
INGRESOS	5	2.33431042	0.46686208	600.79	0.0001	
DENSIDAD	3	0.77587292	0.25862431	332.81	0.0001	
INGRESOS*DENSIDAD	15	0.03201458	0.00213431	2.75	0.0133	



Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Ingresos, lineal	1	0.34750446	0.34750446	447.19	0.0001
Ingresos, cuadratica	1	1.43098229	1.43098229	1841.48	0.0001
Ingresos, cubica	1	0.00646007	0.00646007	8.31	0.0082
Densidad, lineal	1	0.77407042	0.77407042	996.12	0.0001
Densidad, cuadratica	1	0.00025208	0.00025208	0.32	0.5743
Densidad, cubica	1	0.00155042	0.00155042	2.00	0.1706
lineal*lineal	1	0.00003432	0.00003432	0.04	0.8353
lineal*cuadratica	1	0.00007875	0.00007875	0.10	0.7530
cuadratic*lineal	1	0.01348003	0.01348003	17.35	0.0003

### Experimentos para el estudio de las *Superficies de Respuesta*.-

A menudo, clasificaciones y experimentos factoriales de  $n$  vías, en los cuales los niveles de los factores se refieren a cantidades medidas de sendos tratamientos, es decir, todos los factores representan variables cuantitativas, tales como tiempo, temperatura, cantidad de insecticida o cantidad de un componente dietético, pueden considerarse como experimentos *planeados* para determinar la naturaleza a una *superficie de respuesta*. En particular, se puede estar buscando un valor igual o cercano a un máximo, ya que la mayoría de las respuestas no tienen porque ser lineales. Los efectos principales y las interacciones pueden interpretarse fácilmente en términos de tales superficies.

Si denominamos a los rendimientos o respuestas como  $Y$ , tendremos que esta será función de los niveles de las variables que hacen de factores, de esta manera

$$Y_u = \phi(X_{1u}, X_{2u}, \dots, X_{ku}) + e_u$$

donde  $u= 1, 2, \dots, N$  que representa las  $N$  observaciones en el experimento factorial y  $X_{iu}$  representa el nivel del  $i$ -ésimo factor en la  $u$ -ésima observación. A la función  $\phi$  se le denomina *superficie de respuesta*. El residuo  $e_u$  mide el error experimental de la  $u$ -ésima observación. Un conocimiento de la función  $\phi$  da un resumen completo de los resultados del experimento y permite predecir las respuestas para valores de  $X_{iu}$  que no fueron probados en el experimento.

Cuando no se conoce la forma lineal de  $\phi$ , la función puede aproximarse satisfactoriamente dentro de la región experimental por un polinomio de las variables  $X_{iu}$

Estos experimentos son muy útiles en trabajos de investigación y desarrollo en la industria donde el experimentador está buscando, por ejemplo, los medios de elaborar un producto nuevo con ciertas características deseables u obtener más económicamente un producto. También se encuentran ejemplos en otros campos, como puede ser el estudio del mejor medio de cultivo para cierto organismo de laboratorio.

La respuesta puede ser la cantidad real de producto o una medida de la calidad del mismo. Algunas veces, el objetivo es minimizar  $Y$ , si esta representa la cantidad de un subproducto indeseable o el costo de producción por unidad.

Además de localizar el máximo (o mínimo) de  $Y$ , es útil aprender algo de cómo varía  $Y$  en la vecindad del máximo, cuando los niveles de los factores se cambian de sus valores óptimos. Esto es así por varias razones.

- 1) Si se aplican a gran escala los resultados, puede no ser factible situar cada factor exactamente a su nivel óptimo. Alguna combinación de niveles diferente del óptimo puede ser más económica de mantener.
- 2) Frecuentemente el producto tendrá más de una característica importante. Un cambio en la  $x_i$ , alejándola del óptimo, puede ser preferible a causa de su efecto sobre una de las otras características.
- 3) La forma de la superficie de respuesta cerca del óptimo puede dar indicios de la naturaleza del proceso.
- 4) La superficie puede carecer de un máximo verdadero en la región de experimentación (o en la región en la cual el proceso se está estudiando). En este caso, se desea conocer la naturaleza de la superficie en zonas de respuesta relativamente alta.

Para continuar la explicación, hagámoslo con un ejemplo.

### Ejemplo.-

Se está estudiando la pérdida de ácido ascórbico, conforme transcurre el tiempo, en judías almacenadas a tres temperaturas. Las muestras fueron tomadas de manera uniforme antes de las ocho de la mañana y congeladas inmediatamente. De entre todas las unidades experimentales muestreadas se tomaron al azar 3 y se almacenaron durante dos semanas a  $-18^{\circ}\text{C}$  ( $0^{\circ}\text{F}$ ), otra tres se almacenaron a la misma temperatura dos semanas más que las anteriores (cuatro semanas), y así hasta ocho semanas. Las otras dos temperaturas de almacenamiento fueron de  $12^{\circ}\text{C}$  ( $10^{\circ}\text{F}$ ) y  $-7^{\circ}\text{C}$  ( $20^{\circ}\text{F}$ ). En todos los tratamientos se hicieron tres repeticiones de una manera completamente aleatoria. Los datos son

		Semanas de almacenamiento				$\Sigma$
		2	4	6	8	
Temperatura	0	13	14	13	14	184
		15	16	15	15	
		17	17	18	17	
De	10	14	12	12	10	166
		15	14	14	12	
		16	17	15	15	
Almacenamiento	20	9	8	6	3	99
		11	9	7	5	
		14	11	8	8	
$\Sigma$		124	118	108	99	449

Como se ve en los totales la concentración de ácido ascórbico disminuye cuando aumenta la temperatura de almacenamiento y también es inversa la relación

con el tiempo de almacenamiento. Si se miran los totales dentro de temperaturas y de tiempo de almacenamiento se observa la misma relación inversa, excepto para los 0°F. Una observación más detallada podría sugerir que la disminución del ácido ascórbico con el aumento de temperatura de almacenamiento no es lineal y no tiene la misma pendiente en los diferentes periodos de almacenamiento. El análisis de todas estas sugerencias intuitivas es

**Archivo del programa SAS (C10-6.SAS).-**

```
Title 'Superficie de respuesta de semana y temperatura';
option ls=75 ps=60;
data spfprt;
infile 'c10-6.dat';
input semana tempera ascorbi @@;
proc glm;
class tempera semana;
model ascorbi = semana tempera tempera*semana ;
contrast 'Temperatura lineal' tempera -1 0 1;
contrast 'Temperatura cuadra' tempera 1 -2 1;
contrast 'Semana lineal' semana -3 -1 1 3;
contrast 'Semana cuadra' semana 1 -1 -1 1;
contrast 'Semana cubica' semana -1 3 -3 1;
contrast 'T-lineal*S-lineal' tempera*semana 3 1 -1 -3 0 0 0 0 -3 -1 1 3;
contrast 'T-lineal*S-cuadra' tempera*semana -1 1 1 -1 0 0 0 0 1 -1 -1 1;
contrast 'T-lineal*S-cubica' tempera*semana 1 -3 3 -1 0 0 0 0 -1 3 -3 1;
contrast 'T-cuadra*S-lineal' tempera*semana -3 -1 1 3 6 2 -2 -6 -3 -1 1 3;
contrast 'T-cuadra*S-cuadra' tempera*semana 1 -1 -1 1 -2 2 2 -2 1 -1 -1 1;
contrast 'T-cuadra*S-cubica' tempera*semana -1 3 -3 1 2 -6 6 -2 -1 3 -3 1;
run;
```

**Archivo de datos (C10-6.DAT).-**

2	0	13	4	0	14	6	0	13	8	0	14
2	0	15	4	0	16	6	0	15	8	0	15
2	0	17	4	0	17	6	0	18	8	0	17
2	10	14	4	10	12	6	10	12	8	10	10
2	10	15	4	10	14	6	10	14	8	10	12
2	10	16	4	10	17	6	10	15	8	10	15
2	20	9	4	20	8	6	20	6	8	20	3
2	20	11	4	20	9	6	20	7	8	20	5
2	20	14	4	20	11	6	20	8	8	20	8

**Archivo de resultados (C10-6.LST).-**

Dependent Variable: ASCORBI						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	11	408.972222	37.179293	9.49	0.0001	
Error	24	94.000000	3.916667			
Corrected Total	35	502.972222				
	R-Square	C.V.	Root MSE	ASCORBI Mean		
	0.813111	15.86772	1.97906	12.4722		

Source	DF	Type I SS	Mean Square	F Value	Pr > F
SEMANA	3	40.527778	13.509259	3.45	0.0324
TEMPERA	2	334.388889	167.194444	42.69	0.0001
TEMPERA*SEMANA	6	34.055556	5.675926	1.45	0.2376

Source	DF	Type III SS	Mean Square	F Value	Pr > F
SEMANA	3	40.527778	13.509259	3.45	0.0324
TEMPERA	2	334.388889	167.194444	42.69	0.0001
TEMPERA*SEMANA	6	34.055556	5.675926	1.45	0.2376

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Temperatura lineal	1	301.041667	301.041667	76.86	0.0001
Temperatura cuadrada	1	33.347222	33.347222	8.51	0.0075
Semana lineal	1	40.138889	40.138889	10.25	0.0038
Semana cuadrada	1	0.250000	0.250000	0.06	0.8027
Semana cubica	1	0.138889	0.138889	0.04	0.8522
T-lineal*S-lineal	1	33.075000	33.075000	8.44	0.0078
T-lineal*S-cuadra	1	0.375000	0.375000	0.10	0.7597
T-lineal*S-cubica	1	0.008333	0.008333	0.00	0.9636
T-cuadra*S-lineal	1	0.136111	0.136111	0.03	0.8537
T-cuadra*S-cuadra	1	0.125000	0.125000	0.03	0.8597
T-cuadra*S-cubica	1	0.336111	0.336111	0.09	0.7721

Se observa que con respecto a la temperatura, la regresión cuadrática es significativa, por lo que la línea podría ser una parábola, es decir, que la calidad de este producto decrece con una rapidez acelerada conforme aumenta la temperatura de almacenamiento. Mientras que con respecto a las semanas de almacenamiento, solamente da significativa la relación lineal, es decir que la disminución de la calidad es uniforme a lo largo de las semanas de almacenamientos pues sigue una recta.

Por tanto se tiene que son significativas las pendiente negativas de la temperatura para la recta y la parábola, mientras que solamente la recta es significativa para las semanas, por lo que la pregunta lógica es de qué manera cambia la regresión lineal de la temperatura conforme pasan las semanas de almacenamiento. Para ello se elaboran las interacciones de las componentes lineal y Cuadrática de la temperatura con las semanas de almacenamiento, dando significativa solamente la interacción linealxlineal.

Estos resultados pueden expresarse como una relación matemática entre el ácido ascórbico  $Y$ , la temperatura de almacenamiento  $T$  y las semanas de almacenamiento  $S$ . Para describir la relación adecuadamente se necesitan los términos de la temperatura lineal ( $T_L$ ), temperatura cuadrática ( $T_C$ ), semana lineal ( $S_L$ ) y la interacción temperatura lineal por semana lineal ( $T_L S_L$ ), por lo que se necesitarán los coeficientes polinomiales de estas relaciones para cada una de las doce combinaciones de tratamientos, es decir

Tratamientos		Y	X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>	X <sub>4</sub>	Ŷ	
Temp	Sema	Σ	T <sub>L</sub>	T <sub>C</sub>	S <sub>L</sub>	T <sub>L</sub> S <sub>L</sub>		
0	2	45	-1	1	-3	3	45.53	
0	4	47	-1	1	-1	1	45.84	
0	6	46	-1	1	1	-1	46.16	
0	8	46	-1	1	3	-3	46.47	
10	2	45	0	-2	-3	0	45.75	
10	4	43	0	-2	-1	0	42.92	
10	6	41	0	-2	1	0	40.08	
10	8	37	0	-2	3	0	37.25	
20	2	34	1	1	-3	-3	33.73	
20	4	28	1	1	-1	-1	27.74	
20	6	21	1	1	1	1	21.76	
20	8	16	1	1	3	3	15.77	
		Σ=-449	ΣX <sub>i</sub> Y	-85	-49	-85	-63	
			ΣX <sub>i</sub> <sup>2</sup>	8	24	60	40	

La función lineal será de la forma

$$\hat{Y} = \bar{Y} + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4$$

donde  $\hat{Y}$  es el total de ácido ascórbico previsto en cada casilla por la ecuación;  $X_1$  es el coeficiente polinomial correspondiente a la relación lineal de la temperatura;  $X_2$  es el coeficiente polinomial correspondiente a la relación cuadrática de la temperatura;  $X_3$  es el coeficiente polinomial correspondiente a la relación lineal de las semanas de almacenamiento; y  $X_4$  es el coeficiente polinomial correspondiente a la relación T-lineal x S-lineal. Los coeficientes de regresión son

$$b_i = \frac{\sum X_i Y}{\sum X_i^2}$$

siendo  $X_i$  los valores correspondientes de los coeficientes polinomiales, e  $Y$  los valores totales de cada casilla/tratamiento, tal como se muestran en la tabla anterior.

Si se calculan los coeficientes de la ecuación anterior se obtienen

$$\bar{Y} = \frac{449}{12} = 37.4167$$

$$b_1 = \frac{-85}{8} = -10.625$$

$$b_2 = \frac{-49}{24} = -2.0417$$

$$b_3 = \frac{-85}{60} = -1.4167$$

$$b_4 = \frac{-63}{40} = -1.575$$

Por lo que se tiene la ecuación

$$\hat{Y} = 37.417 - 10.625X_1 - 2.042X_2 - 1.417X_3 - 1.575X_4$$

Como los valores de  $X_i$  son los de los coeficientes de los polinomios, expresados en la tabla anterior, es fácil el cálculo de los valores estimados de ácido ascórbico para cada combinación de tratamientos. Por ejemplo, para  $10^\circ F$  y 4 semanas de almacenamiento

$$\hat{Y}_{10-4} = 37.417 - 10.625(0) - 2.042(-2) - 1.417(-1) - 1.575(0) = 42.918$$

En la tabla anterior se mostraron estas cantidades estimadas.

Si se quiere expresar la ecuación de predicción en términos de la *Temperatura* y las *Semanas* de almacenamiento, conocido ya los términos significativos se pueden ajustar los datos a dicha función con el siguiente programa SAS.

#### Archivo del programa SAS (C10-7.SAS).-

```

title 'Ecuación de predicción';
option ls=75 ps=60;
data spfprt;
infile 'C10-6.dat';
input semana tempera ascorbi @@;
proc glm;
model ascorbi = tempera tempera*tempera semana
tempera*semana;
run;

```

#### Archivo de resultados (C10-7.LST).-

Dependent Variable: ASCORBI					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	407.602778	101.900694	33.12	0.0001
Error	31	95.369444	3.076434		
Corrected Total	35	502.972222			

	R-Square	C.V.	Root MSE	ASCORBI Mean	
	0.810388	14.06306	1.75398	12.4722	
Source	DF	Type I SS	Mean Square	F Value	Pr > F
TEMPERA	1	301.041667	301.041667	97.85	0.0001
TEMPERA*TEMPERA	1	33.347222	33.347222	10.84	0.0025
SEMANA	1	40.138889	40.138889	13.05	0.0011
TEMPERA*SEMANA	1	33.075000	33.075000	10.75	0.0026
Source	DF	Type III SS	Mean Square	F Value	Pr > F
TEMPERA	1	13.3703704	13.3703704	4.35	0.0454
TEMPERA*TEMPERA	1	33.3472222	33.3472222	10.84	0.0025
SEMANA	1	0.2005556	0.2005556	0.07	0.8002
TEMPERA*SEMANA	1	33.0750000	33.0750000	10.75	0.0026
Parameter	Estimate	T for H0: Parameter=0	Pr >  T	Std Error of Estimate	
INTERCEPT	15.06944444	13.09	0.0001	1.15090209	
TEMPERA	0.31666667	2.08	0.0454	0.15189882	
TEMPERA*TEMPERA	-0.02041667	-3.29	0.0025	0.00620124	
SEMANA	0.05277778	0.26	0.8002	0.20670812	
TEMPERA*SEMANA	-0.05250000	-3.28	0.0026	0.01601154	

De manera que la ecuación de predicción de la cantidad media de ácido ascórbico por combinación de tratamientos es

$$\hat{Y} = 15.0694 + 0.3167T - 0.0204T^2 + 0.0528S - 0.0525TS$$

Es decir, que la cantidad de ácido ascórbico que se espera que quede en este producto después de 6 semanas de almacenamiento a 10°F es

$$\hat{Y} = 15.0694 + 0.3167 \times 10 - 0.0204 \times 10^2 + 0.0528 \times 6 - 0.0525 \times 10 \times 6 = 13.3632$$

Esta relación es la que se conoce como de *Superficie Respuesta*. Con factores cuantitativos, se puede investigar cuál es la combinación de tratamientos que propician una mayor (o una menor) respuesta de la variable medida. Por ejemplo, en el caso del ácido ascórbico del ejemplo anterior, la máxima respuesta (mayor cantidad de ácido ascórbico conservado) se consigue a 5.14°F y 2.4 semanas de almacenamientos; y la mínima respuesta se consigue a 19.34°F y 6.1 semanas de almacenamiento.

El paquete estadístico SAS posee un procedimiento para análisis de superficie respuesta, por lo que todo el desarrollo anterior, que se ha hecho con fines de comprensión, se podría haber hecho con el siguiente programa

**Archivo del programa SAS (C10-8.SAS).-**

```
Title 'Experimento Superficie Respuesta';
option ls=75 ps=60;;
data spfprt;
infile 'c10-6.dat';
input semana tempera ascorbi @@;
proc sort;
  by semana tempera;
run;
proc rsreg;
model ascorbi = semana tempera / lackfit;
ridge min;
run;
proc rsreg;
model ascorbi = semana tempera / lackfit;
ridge max;
run;
```

**Archivo de resultados (C10-8.LST).-**

Coding Coefficients for the Independent Variables					
Factor	Subtracted off	Divided by			
SEMANA	5.000000	3.000000			
TEMPERA	10.000000	10.000000			

Response Surface for Variable ASCORBI					
Response Mean	12.472222				
Root MSE	1.780631				
R-Square	0.8109				
Coef. of Variation	14.2768				

Degrees of Freedom					
Regression	Freedom	Type I Sum of Squares	R-Square	F-Ratio	Prob > F
Linear	2	341.180556	0.6783	53.803	0.0000
Quadratic	2	33.597222	0.0668	5.298	0.0107
Crossproduct	1	33.075000	0.0658	10.432	0.0030
Total Regress	5	407.852778	0.8109	25.727	0.0000

Degrees of Freedom					
Residual	Freedom	Sum of Squares	Mean Square	F-Ratio	Prob > F
Lack of Fit	6	1.119444	0.186574	0.0476	0.9995
Pure Error	24	94.000000	3.916667		
Total Error	30	95.119444	3.170648		

Degrees of Freedom					
Parameter	Freedom	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob >  T
INTERCEPT	1	14.652778	1.888645	7.758	0.0000
SEMANA	1	0.261111	0.771036	0.339	0.7372
TEMPERA	1	0.316667	0.154207	2.054	0.0488
SEMANA*SEMANA	1	-0.020833	0.074193	-0.281	0.7808
TEMPERA*SEMANA	1	-0.052500	0.016255	-3.230	0.0030
TEMPERA*TEMPERA	1	-0.020417	0.006295	-3.243	0.0029



Parameter	Parameter Estimate from Coded Data
INTERCEPT	13.937500
SEMANA	-1.416667
TEMPERA	-3.541667
SEMANA*SEMANA	-0.187500
TEMPERA*SEMANA	-1.575000
TEMPERA*TEMPERA	-2.041667

Factor	Degrees of Freedom	Sum of Squares	Mean Square	F-Ratio	Prob > F
SEMANA	3	73.463889	24.487963	7.723	0.0006
TEMPERA	3	367.463889	122.487963	38.632	0.0000

Canonical Analysis of Response Surface  
(based on coded data)

Factor	Critical Value	
	Coded	Uncoded
SEMANA	0.217614	5.652842
TEMPERA	-0.951284	0.487163

Predicted value at stationary point      15.467922

Eigenvectors

Eigenvalues	SEMANA	TEMPERA
0.101821	0.938656	-0.344855
-2.330988	0.344855	0.938656

Stationary point is a saddle point.

Estimated Ridge of Minimum Response for Variable ASCORBI

Coded Radius	Estimated Response	Standard Error	Uncoded Factor Values	
			SEMANA	TEMPERA
0.0	13.937500	0.633905	5.000000	10.000000
0.1	13.532758	0.630861	5.110521	10.929666
0.2	13.081425	0.621959	5.219613	11.861212
0.3	12.583491	0.607927	5.327669	12.794106
0.4	12.038953	0.590101	5.434951	13.727998
0.5	11.447807	0.570598	5.541639	14.662650
0.6	10.810050	0.552533	5.647861	15.597892
0.7	10.125680	0.540177	5.753712	16.533604
0.8	9.394697	0.538826	5.859260	17.469695
0.9	8.617098	0.554061	5.964561	18.406097
1.0	7.792884	0.590442	6.069656	19.342755

Estimated Ridge of Maximum Response for Variable ASCORBI

Coded Radius	Estimated Response	Standard Error	Uncoded Factor Values	
			SEMANA	TEMPERA
0.0	13.937500	0.633905	5.000000	10.000000
0.1	14.295662	0.630864	4.887427	9.073073
0.2	14.607264	0.621975	4.771760	8.150355
0.3	14.872345	0.607960	4.651111	7.234586
0.4	15.090979	0.590096	4.521680	6.331500
0.5	15.263348	0.570281	4.374623	5.455279

0.6	15.390001	0.550867	4.184568	4.650991
0.7	15.473382	0.533643	3.865623	4.109151
0.8	15.525179	0.526032	3.338040	4.228528
0.9	15.565163	0.550487	2.823332	4.674887
1.0	15.602440	0.613132	2.378577	5.137284

**Comparación múltiple de medias o contrastes a posteriori o pruebas sugeridas por los datos.-**

Este tipo de comparación se realiza cuando no se encuentra un contraste lógico entre los diferentes niveles de los tratamientos o cuando es ese precisamente nuestro objetivo, es decir, comparar todos los posibles pares de medias.

Las pruebas que se van a describir a continuación pueden realizarse en un diseño con más de un factor, para los diferentes niveles uno de los factores, teniendo como única precaución la de utilizar el  $CM_{Error}$  del ANOVA realizado con todos los factores.

Para realizar la comparación entre las diferentes medias, primero se ordenan dichas medias de mayor a menor, por ejemplo

$$\bar{X}_4 > \bar{X}_3 > \bar{X}_1 > \bar{X}_2 > \bar{X}_5 > \bar{X}_6$$

Después se hace un cuadro de doble entrada en el que se ordenan las medias de mayor a menor en vertical y de menor a mayor en horizontal y se realizan las diferencias disponiéndolas en triángulo, de forma que a medida que se desciende en las columnas o se mueve hacia la derecha en las filas el número de medias comprendidas entre las dos que se comparan (ambas inclusive) disminuye de uno en uno, conforme se expresa en la  $p$  de cada casilla de la tabla siguiente

Medias	$\bar{X}_6$	$\bar{X}_5$	$\bar{X}_2$	$\bar{X}_1$	$\bar{X}_3$
$\bar{X}_4$	$\bar{X}_4 - \bar{X}_6$ $p=6$	$\bar{X}_4 - \bar{X}_5$ $p=5$	$\bar{X}_4 - \bar{X}_2$ $p=4$	$\bar{X}_4 - \bar{X}_1$ $p=3$	$\bar{X}_4 - \bar{X}_3$ $p=2$
$\bar{X}_3$	$\bar{X}_3 - \bar{X}_6$ $p=5$	$\bar{X}_3 - \bar{X}_5$ $p=4$	$\bar{X}_3 - \bar{X}_2$ $p=3$	$\bar{X}_3 - \bar{X}_1$ $p=2$	
$\bar{X}_1$	$\bar{X}_1 - \bar{X}_6$ $p=4$	$\bar{X}_1 - \bar{X}_5$ $p=3$	$\bar{X}_1 - \bar{X}_2$ $p=2$		
$\bar{X}_2$	$\bar{X}_2 - \bar{X}_6$ $p=3$	$\bar{X}_2 - \bar{X}_5$ $p=2$			
$\bar{X}_5$	$\bar{X}_5 - \bar{X}_6$ $p=2$				

Una vez obtenidas todas las posibles diferencias que se pueden formar con las medias, cada una se comparará con el valor crítico que le corresponde según los

diferentes métodos que se verán inmediatamente. Se comienza comparando la media mayor con todas las demás, por lo que se comienza con la primera fila de la tabla anterior. Cuando una diferencia no sea significativa ya no lo serán las diferencias de la parte derecha de su fila, por lo que no hay necesidad de comprobar dichas significaciones. Si se quiere comparar la segunda media con todas las demás, nos fijamos en la segunda fila, etc.

De los varios métodos, descritos por distintos autores, que se podrían emplear para contrastar si estas diferencias son significativas, nos vamos a circunscribir exclusivamente a cinco.

**Prueba LSD (Least-Significant-Difference).-**

También llamada prueba *t múltiple*. Se puede usar, también, para pruebas a priori, aunque se utiliza con más frecuencia para pruebas sugeridas por los datos. Para el cálculo de la región crítica se usa la tabla *t* de la misma forma que se expuso para contrastar dos medias, aplicándolo repetidamente a todas las parejas en las que estamos interesados. Es decir, comprobando, al nivel  $\alpha$  que nos impongamos, si las diferencias de medias tomadas dos a dos cumplen que

$$\bar{X}_i - \bar{X}_j > S_{\bar{X}} \sqrt{2} t_{(gl_{error}; \alpha/2)}$$

donde  $S_{\bar{X}}$  es el error típico combinado, es decir, la raíz cuadrada del cuadrado medio del error dividido por el tamaño de submuestra. Esto es

$$S_{\bar{X}} = \sqrt{\frac{CM_{Error}}{n}}$$

En el caso de que los tamaños de las submuestras sean diferentes (experimento desequilibrado), se utiliza como  $n$  el valor de la media armónica de las  $t$  submuestras. El contraste sería entonces

$$\bar{X}_i - \bar{X}_j > \sqrt{CM_{Error} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)} t_{(gl_{error}; \alpha/2)}$$

**Prueba Tukey o de recorrido Studentizado.-**

Utiliza la tabla  $q$  (Tabla 8) para el cálculo de la región crítica y consiste como en la anterior, en contrastar las parejas en las que se esta interesado a base de ver si las diferencias son significativas, según la expresión

$$\bar{X}_i - \bar{X}_j > S_{\bar{X}} q_{(gl_{error}; \alpha/2; p_{medias})}$$

Siendo

$S_{\bar{x}}$  el error típico combinado, es decir, la raíz cuadrada del cuadrado medio del error dividido por el tamaño de submuestra.

$q$  valores de la tabla 8 para  $p$ =número total de medias y  $gl$  del error.

No se incluye el término  $\sqrt{2}$  porque va incluido en los valores de  $q$  de la tabla.

Considérese, así mismo, que si en la  $t$  solamente hay que considerar el  $\alpha$  establecido y los  $gl$ , en esta prueba hay que tener en cuenta, también, el número total de medias que se van a contrastar, esto es el valor de  $p$ .

En el caso de tamaños de submuestras diferentes se utilizaría como  $n$  el valor de la media armónica de las dos submuestras, siendo la prueba

$$\bar{X}_i - \bar{X}_j > \sqrt{CM_{Error} \frac{1}{2} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)} q_{(gl_{error}; \alpha/2; P_{medias})}$$

### Prueba de *Student-Newmans-Keuls* o *SNK*.-

Se basa en la tabla  $q$  (tabla 8), de recorrido *Studentizado*, pero en lugar de tomar un solo valor para la región crítica, correspondiente a la  $p$  del número total de medias que se van a comparar, como se hizo en la prueba anterior, la región crítica cambiará, dentro de la misma prueba, con arreglo a la distancia, en número de medias, entre las dos medias que se comparan. Por tanto, los valores de  $p$  dentro de una misma prueba cambiarán de un mínimo de  $p=2$ , correspondiente a dos medias cercanas en valor (sin otra media entre el valor de ambas), hasta un valor máximo de  $p=t-1$ , correspondiente a las dos medias de valor más alejado.

Por lo que para cada valor de  $q$  se calcula su región crítica multiplicando el valor  $q$  de la tabla por el error típico combinado de las medias, quedando la prueba de la siguiente manera

$$\bar{X}_i - \bar{X}_j > S_{\bar{x}} q_{(gl_{error}; \alpha/2; P_{medias})}$$

Siendo

$S_{\bar{x}}$  el error típico combinado, es decir, la raíz cuadrada del cuadrado medio del error dividido por el tamaño de submuestra.

$q$  valores de la tabla 8 para  $p$ =número de medias entre las dos que se están contrastando y  $gl$  del error.

En el caso de tamaños de submuestras diferentes se utilizaría como  $n$  el valor de la media armónica de las dos submuestras, siendo la prueba

$$\bar{X}_i - \bar{X}_j > \sqrt{CM_{Error} \frac{1}{2} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)} q_{(gl_{error}; \alpha/2; p_{medias})}$$

### Prueba de *Duncan* o de amplitudes múltiples.-

Se parece a la prueba *SNK* en que usa amplitudes múltiples y regiones críticas variables que dependen del número de medias que entran en cada etapa. Se usa la tabla *r* (Tabla 7) para los rangos críticos.

Para efectuar las comparaciones múltiples entre *t* medias, se necesita, como en las anteriores pruebas, el cuadrado medio del error sus grados de libertad y el número de observaciones (*n*) en cada nivel del factor.

El contraste consiste en

$$\bar{X}_i - \bar{X}_j > S\bar{X} r_{(gl_{error}; \alpha/2; p_{medias})}$$

Siendo

$S\bar{X}$  el error típico combinado, es decir, la raíz cuadrada del cuadrado medio del error dividido por el tamaño de submuestra.

*r* valores de la tabla 7 para *p*=número de medias entre las dos que se están contrastando y *gl* del error.

En el caso de tamaños de submuestras diferentes se utilizaría como *n* el valor de la media armónica de las dos submuestras, siendo la prueba

$$\bar{X}_i - \bar{X}_j > \sqrt{CM_{Error} \frac{1}{2} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)} r_{(gl_{error}; \alpha/2; p_{medias})}$$

### Prueba de *Scheffe*.-

Es muy general en el sentido de que todas las posibles comparaciones pueden probarse en cuanto a significación, es decir que no solamente se pueden establecer contrastes entre dos medias sino entre ciertas combinaciones lineales de ellas, no siendo necesario que el número de elementos por tratamiento sea igual para todos ellos. El contraste consiste en

$$\bar{X}_i - \bar{X}_j > S\bar{X} \sqrt{2} F_1$$

siendo

$$F_1 = \sqrt{(t-1) F_{(t-1, N-t; \alpha)}}$$

En el caso de tamaños de submuestras diferentes se utilizaría como  $n$  el valor de la media armónica de las dos submuestras, siendo la prueba

$$\bar{X}_i - \bar{X}_j > \sqrt{CM_{Error} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)} F_1$$

**Ejemplo.-**

Una empresa tiene en un establecimiento cinco vendedores y proyecta organizar unas primas en función de la mayor o menos capacidad de venta de cada uno. Una vez contrastado que la capacidad de los vendedores difiere significativamente de unos a otros, con un nivel de significación de 0.05, se desea conocer cuál es el orden de eficacia. Para ello se registra en cinco instantes distintos de tiempo el volumen de ventas realizado por cada uno; estas cantidades en miles de pesetas son

Vendedor 1	Vendedor 2	Vendedor 3	Vendedor 4	Vendedor 5
9.34	6.46	5.79	8.37	4.94
8.53	4.83	5.13	7.57	4.11
9.43	5.89	6.17	8.69	5.45
8.37	5.30	4.72	8.06	5.21
9.64	6.33	5.60	7.23	5.00

El análisis es

<i>FV</i>	<i>gl</i>	<i>SC</i>	<i>CM</i>	<i>F</i>
Vendedores	4	63.3364	15.8341	45.749***
Error	20	6.9222	0.3461	
<i>Total</i>	24	5893		

Al ser  $F_{(4,20; 0.001)} = 7.10$  la  $F_o$  es altamente significativa por lo que se acepta que existe diferencias de eficacia entre los vendedores.

Para establecer el orden de eficacia se actúa tal como se ha indicado en el párrafo anterior.

Primeramente se calcula las cinco medias y las clasificamos de mayor a menor para a continuación realizar el cuadro de diferencias de las medias

$$\bar{X}_1 = 9.06 > \bar{X}_4 = 7.98 > \bar{X}_2 = 5.76 > \bar{X}_3 = 5.48 > \bar{X}_5 = 4.94$$

<i>Medias</i>	$\bar{X}_5$ 4.94	$\bar{X}_3$ 5.48	$\bar{X}_2$ 5.76	$\bar{X}_4$ 7.98
$\bar{X}_1$ 9.06	4.12 $p=5$	3.58 $p=4$	3.30 $p=3$	1.08 $p=2$
$\bar{X}_4$ 7.98	3.04 $p=4$	2.50 $p=3$	2.22 $p=2$	
$\bar{X}_2$ 5.76	0.82 $p=3$	0.28 $p=2$		
$\bar{X}_3$ 5.48	0.54 $p=2$			

### Prueba Duncan.-

Para el cálculo de los rangos críticos ( $R$ ) se toman los valores de  $r$  (Tabla 7) con el recorrido  $p$  y  $gl$  del error

$$R_p = S_{\bar{X}} r_p = \sqrt{\frac{0.3461}{5}} r_p = 0.2631 r_p$$

Los valores  $r$  se toman de la Tabla 7 con el nivel de significación que se considere adecuado (por ejemplo,  $\alpha=0.05$ ) con los  $gl$  del error ( $gl=20$ ) y en la columna del recorrido del número de medias (para medias contiguas  $p=2$ , para medias que disten dos lugares  $p=3$ , etc.). De manera que se obtiene de la Tabla 7 los siguientes valores

$$r_5 = 3.25 \quad r_4 = 3.18 \quad r_3 = 3.10 \quad r_2 = 2.95$$

con lo que los recorridos o rangos críticos son

$$R_5 = 0.8551 \quad R_4 = 0.8367 \quad R_3 = 0.8156 \quad R_2 = 0.7761$$

Obsérvese que el valor crítico mínimo ( $R_2$ ) es el único que se utiliza en la prueba de *LSD* y coincide con el valor crítico mínimo de la prueba *SNK*. Mientras que el valor crítico máximo  $R_5$  es menor que el valor crítico máximo de la prueba *SNK* y menor que el único valor crítico de la prueba de *Tukey* y de la prueba de *Scheffe*.

Como el contraste para saber si las medias  $X_i$  y  $X_j$  son distintas es comparar su diferencia ( $X_i - X_j$ ) con el valor de  $R_p$  correspondiente al recorrido entre  $X_i$  y  $X_j$ , de manera que si la diferencia es mayor que  $R_p$  es porque ambas medias son significativamente diferentes, podemos poner los valores de  $R_p$  en la misma tabla, debajo de las diferencias de las medias, de esta manera

Medias	$\bar{X}_5$	$\bar{X}_3$	$\bar{X}_2$	$\bar{X}_4$
	4.94	5.48	5.76	7.98
$\bar{X}_1$	4.12	3.58	3.30	1.08
9.06	0.8551	0.8367	0.8156	0.7761
$\bar{X}_4$	3.04	2.50	2.22	
7.98	0.8367	0.8156	0.7761	
$\bar{X}_2$	0.82	0.28		
5.76	0.8156	0.7761		
$\bar{X}_3$	0.54			
5.48	0.7761			

Con este resultado se llega a la conclusiones

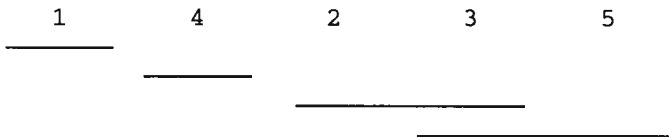
La media  $X_1$  es significativamente superior a todas las demás

La media  $X_4$  es significativamente superior a las que son menores a ella

La media  $X_2$  es mayor que  $X_5$  e igual que  $X_3$  y

Las medias  $X_3$  y  $X_5$  son estadísticamente iguales.

Esto se puede esquematizar así



En realidad, la diferencia  $X_2 - X_5 = 0.82$  es escasamente mayor que  $R_3 = 0.8156$ , pero dado el orden en que se encuentran (2, 3, 5) es más lógico suponer que de hecho son iguales, máxime cuando 0.82 y 0.8156 son dos magnitudes prácticamente iguales.

Por lo que se puede concluir que el mejor vendedor es el n° 1 seguido del n° 4 y los tres restantes son iguales, estando en tercer lugar por volumen de ventas.

### Prueba Student-Newman-Keuls (SNK).-

Conceptualmente es como la anterior sólo que para el cálculo de los rangos críticos ( $Q$ ) se toman los valores de  $q$  (Tabla 8) con el recorrido  $p$  y  $gl$  del error

$$Q_p = S_{\bar{X}} q_p = \sqrt{\frac{0.3461}{5}} q_p = 0.2631 q_p$$



los valores  $q$  se toman de la Tabla 8 con el nivel de significación que se considere adecuado (por ejemplo  $\alpha=0.05$ ) con los  $gl$  del error ( $gl=20$ ) y en la columna del recorrido del número de medias (para medias contiguas  $p=2$ , para medias que disten dos lugares  $p=3$ , etc.). De manera que se obtienen los valores

$$q_5 = 4.23 \quad q_4 = 3.96 \quad q_3 = 3.58 \quad q_2 = 2.95$$

con lo que los recorridos o *rangos críticos* son

$$Q_5 = 1.1129 \quad Q_4 = 1.0419 \quad Q_3 = 0.9419 \quad Q_2 = 0.7761$$

Obsérvese que el valor crítico mínimo ( $Q_2$ ) es el único que se utiliza en la prueba de *LSD* y coincide con el valor crítico mínimo de la prueba *Duncan*. Mientras que el valor crítico máximo  $Q_5$  es el único valor crítico de la prueba de *Tukey* y es muy semejante al único valor crítico de la prueba de *Scheffe*, siendo mayor que el valor crítico máximo de la prueba *Duncan*.

Como el contraste para saber si las medias  $X_i$  y  $X_j$  son distintas es comparar su diferencia ( $X_i - X_j$ ) con el valor de  $Q_p$  correspondiente al recorrido entre  $X_i$  y  $X_j$ , de manera que si la diferencia es mayor que  $Q_p$  es porque ambas medias son significativamente diferentes, podemos poner los valores de  $Q_p$  en la misma tabla, debajo de las diferencias de las medias, de esta manera

<i>Medias</i>	$\bar{X}_5$ 4.94	$\bar{X}_3$ 5.48	$\bar{X}_2$ 5.76	$\bar{X}_4$ 7.98
$\bar{X}_1$ 9.06	4.12 1.1129	3.58 1.0419	3.30 0.9419	1.08 0.7761
$\bar{X}_4$ 7.98	3.04 1.0419	2.50 0.9419	2.22 0.7761	
$\bar{X}_2$ 5.76	0.82 0.9419	0.28 0.7761		
$\bar{X}_3$ 5.48	0.54 0.7761			

Con este resultado se llega a las mismas conclusiones que con la prueba *Duncan*, esto es

La media  $X_1$  es significativamente superior a todas las demás

La media  $X_4$  es significativamente superior a las que son menores a ella

Las tres medias menores son no son significativamente distintas

Esto se puede esquematizar así



Por lo que se puede concluir que el mejor vendedor es el n° 1 seguido del n° 4 y los tres restantes son iguales, estando en tercer lugar por volumen de ventas.

**Prueba Tukey.-**

Es como la anterior solo que no es secuencial y utiliza un solo *rango crítico*, que es el correspondiente a la  $p$  del número total de medias. Para el cálculo de este valor crítico ( $Q$ ) se necesita el valor de  $q$  que se toma de la Tabla 8 con el recorrido  $p$  y  $gl$  del error

$$Q_5 = S_{\bar{X}} q_5 = \sqrt{\frac{0.3461}{5}} q_5 = 0.2631 q_5$$

el valor  $q_5$  se toma de la Tabla 8 con el nivel de significación que se considere adecuado (por ejemplo  $\alpha=0.05$ ) con los  $gl$  del error ( $gl=20$ ) y en la columna del recorrido correspondiente al número total de medias ( $p=5$ ). De manera que se obtiene el valor

$$q_5 = 4.23$$

con lo que el rango crítico es

$$Q_5 = 1.1129$$

Obsérvese que este valor crítico es el valor crítico máximo ( $Q_5$ ) de la prueba *SNK* y es muy semejante al único valor crítico de la prueba de *Scheffe*. Mientras que es mayor que todos los valores críticos de la prueba *Duncan* y mayor que el valor crítico único de la prueba *LSD*.

Como el contraste para saber si las medias  $X_i$  y  $X_j$  son distintas es comparar su diferencia ( $X_i - X_j$ ) con el valor de  $Q_p$  único de manera que si la diferencia es mayor que  $Q_5$  es porque ambas medias son significativamente diferentes, se puede poner este valor de  $Q_5$  en la misma tabla, debajo de todas las diferencias de medias, de esta manera

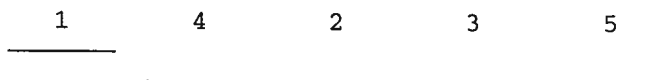
<i>Medias</i>	$\bar{X}_5$ 4.94	$\bar{X}_3$ 5.48	$\bar{X}_2$ 5.76	$\bar{X}_4$ 7.98
$\bar{X}_1$ 9.06	4.12 1.1129	3.58 1.1129	3.30 1.1129	1.08 1.1129
$\bar{X}_4$ 7.98	3.04 1.1129	2.50 1.1129	2.22 1.1129	
$\bar{X}_2$ 5.76	0.82 1.1129	0.28 1.1129		
$\bar{X}_3$ 5.48	0.54 1.1129			

Con este resultado se llega a las conclusiones siguientes

Las medias  $X_1$  y  $X_4$  son significativamente iguales entre ellas y superior a todas las demás

Las tres medias menores son no son significativamente distintas

Esto se puede esquematizar así



Por lo que se puede concluir que los mejores vendedor son el n° 1 y el n° 4, y los tres restantes son iguales y están en segundo lugar en volumen de ventas.

### Prueba **LSD** o de mínima diferencia significativa.-

Es como la anterior en el sentido de que utiliza un sólo valor crítico, que es el correspondiente a la  $t$  de *Student* de manera que para el valor crítico se necesita la  $t$  de la Tabla 2 para el nivel de significación que se crea adecuado y los  $g_l$  del error

$$L_t = S\bar{X}\sqrt{2} \ t_{(20; 0.05/2)} = \sqrt{\frac{0.3461}{5}} \sqrt{2} \ 2.086 = 0.7762$$

Obsérvese que este valor crítico es el valor crítico mínimo de las pruebas *Duncan* y *SNK* ( $R_2$  y  $Q_2$ ) y es menor que el valor crítico único de las pruebas *Tukey* y *Scheffe*.

Como el contraste para saber si las medias  $X_i$  y  $X_j$  son distintas es comparar su diferencia ( $X_i - X_j$ ) con el valor  $L_t$  único, de manera que si la diferencia es mayor que  $L_t$

es porque ambas medias son significativamente diferentes, se puede poner este valor de  $L_1$  en la misma tabla, debajo de todas las diferencias de medias, de esta manera

<i>Medias</i>	$\bar{X}_5$ 4.94	$\bar{X}_3$ 5.48	$\bar{X}_2$ 5.76	$\bar{X}_4$ 7.98
$\bar{X}_1$ 9.06	4.12 0.7762	3.58 0.7762	3.30 0.7762	1.08 0.7762
$\bar{X}_4$ 7.98	3.04 0.7762	2.50 0.7762	2.22 0.7762	
$\bar{X}_2$ 5.76	0.82 0.7762	0.28 0.7762		
$\bar{X}_3$ 5.48	0.54 0.7762			

Con este resultado se llega a las mismas conclusiones que con la prueba *Duncan*, estas son

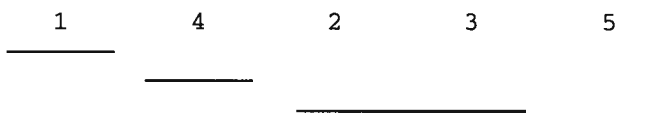
La media  $X_1$  es significativamente superior a todas las demás

La media  $X_4$  es significativamente superior a las que son menores a ella

La media  $X_2$  es mayor que  $X_5$  e igual que  $X_3$  y

Las medias  $X_3$  y  $X_5$  son estadísticamente iguales.

Esto se puede esquematizar así



En realidad, la diferencia  $X_2 - X_5 = 0.82$  es escasamente mayor que  $L = 0.7762$ , pero dado el orden en que se encuentran (2, 3, 5) es más lógico suponer que de hecho son iguales, máxime cuando 0.82 y 0.7762 son dos magnitudes prácticamente iguales.

Por lo que se puede concluir que el mejor vendedor es el n° 1 seguido del n° 4 y los tres restantes son iguales, estando en tercer lugar por volumen de ventas.

### Prueba Scheffe.-

Para el cálculo del *valor crítico* se necesita el valor de la  $F$  (Tabla 3) para los grados de libertad del factor cuyas medias de los diferentes niveles se están

contrastando y los grados de libertad del término de error, y con el nivel de significación que se crea adecuado.

$$F_{(4,20; 0.05)} = 2.87$$

Por lo que se tiene que

$$F_1 = \sqrt{(t-1) F_{(t-1, N-t; \alpha)}} = \sqrt{4 \times 2.87} = 3.3882$$

Obsérvese que este valor crítico es semejante al valor crítico máximo de la prueba *SNK* ( $Q_5$ ) y semejante al único valor crítico de la prueba *Tukey*. Mientras que es mayor que los valores críticos de las pruebas *Duncan* y *LSD*.

Teniendo en cuenta que todas las  $n$  son iguales se va a tener un sólo *valor crítico* con el que se va a comparar todas las diferencias, este valor es

$$L_F = S_{\bar{X}} \sqrt{2} F_1 = \sqrt{\frac{0.3461}{5}} \sqrt{2} 3.3882 = 1.2607$$

Como el contraste para saber si las medias  $X_i$  y  $X_j$  son distintas es comparar su diferencia ( $X_i - X_j$ ) con el valor de  $L_F$  único de manera que si la diferencia es mayor que  $L_F$  es porque ambas medias son significativamente diferentes, se puede poner este valor de  $L_F$  en la misma tabla, debajo de todas las diferencias de medias, de esta manera

<i>Medias</i>	$\bar{X}_5$ 4.94	$\bar{X}_3$ 5.48	$\bar{X}_2$ 5.76	$\bar{X}_4$ 7.98
$\bar{X}_1$ 9.06	4.12 1.2607	3.58 1.2607	3.30 1.2607	1.08 1.2607
$\bar{X}_4$ 7.98	3.04 1.2607	2.50 1.2607	2.22 1.2607	
$\bar{X}_2$ 5.76	0.82 1.2607	0.28 1.2607		
$\bar{X}_3$ 5.48	0.54 1.2607			

Con este resultado se llega a las conclusiones siguientes

Las medias  $X_1$  y  $X_4$  son significativamente iguales entre ellas y superior a todas las demás

Las tres medias menores son no son significativamente distintas

Esto se puede esquematizar así

Por lo que se puede concluir que los mejores vendedor son el nº 1 y el nº 4, y los tres restantes son iguales y están en segundo lugar en volumen de ventas.

**Archivo del programa SAS (C10-9.SAS).-**

```

title 'Comparación múltiple de medias';
options ls=75 ps=60;
data media;
infile 'c10-9.dat';
input vendedor ventas @@;
proc anova;
class vendedor;
model ventas = vendedor;
means vendedor / duncan lsd tukey snk scheffe;
run;
    
```

**Archivo de datos (C10-9.DAT).-**

1	9.34	2	6.46	3	5.79	4	8.37	5	4.94
1	8.53	2	4.83	3	5.13	4	7.57	5	4.11
1	9.43	2	5.89	3	6.17	4	8.69	5	5.45
1	8.37	2	5.30	3	4.72	4	8.06	5	5.21
1	9.64	2	6.33	3	5.60	4	7.23	5	5.00

**Archivo de resultados (C10-9.LST).-**

Dependent Variable: VENTAS					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	63.3363360	15.8340840	45.75	0.0001
Error	20	6.9222400	0.3461120		
Corrected Total	24	70.2585760			
	R-Square	C.V.	Root MSE	VENTAS Mean	
	0.901475	8.851601	0.58831	6.64640	
Source	DF	Anova SS	Mean Square	F Value	Pr > F
VENDEDOR	4	63.3363360	15.8340840	45.75	0.0001
T tests (LSD) for variable: VENTAS					
NOTE: This test controls the type I comparisonwise error rate not the experimentwise error rate.					
Alpha= 0.05 df= 20 MSE= 0.346112					
Critical Value of T= 2.09					
Least Significant Difference= 0.7761					
Means with the same letter are not significantly different.					
T Grouping	Mean	N	VENDEDOR		
A	9.0620	5	1		
B	7.9840	5	4		
C	5.7620	5	2		
D	5.4820	5	3		
D	4.9420	5	5		

Duncan's Multiple Range Test for variable: VENTAS

NOTE: This test controls the type I comparisonwise error rate, not the experimentwise error rate

Alpha= 0.05 df= 20 MSE= 0.346112

Number of Means      2      3      4      5

Critical Range    .7761 .8147 .8392 .8563

Means with the same letter are not significantly different.

Duncan Grouping	Mean	N	VENDEDOR
A	9.0620	5	1
B	7.9840	5	4
C	5.7620	5	2
D      C	5.4820	5	3
D	4.9420	5	5

Student-Newman-Keuls test for variable: VENTAS

NOTE: This test controls the type I experimentwise error rate under the complete null hypothesis but not under partial null hypotheses.

Alpha= 0.05 df= 20 MSE= 0.346112

Number of Means      2      3      4      5

Critical Range    0.7761493 0.9413601 1.0414333 1.1134083

Means with the same letter are not significantly different.

SNK Grouping	Mean	N	VENDEDOR
A	9.0620	5	1
B	7.9840	5	4
C	5.7620	5	2
C	5.4820	5	3
C	4.9420	5	5

Tukey's Studentized Range (HSD) Test for variable: VENTAS

NOTE: This test controls the type I experimentwise error rate, but generally has a higher type II error rate than REGWQ.

Alpha= 0.05 df= 20 MSE= 0.346112

Critical Value of Studentized Range= 4.232

Minimum Significant Difference= 1.1134

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	VENDEDOR
A	9.0620	5	1
A	7.9840	5	4
B	5.7620	5	2
B	5.4820	5	3
B	4.9420	5	5

Scheffe's test for variable: VENTAS

NOTE: This test controls the type I experimentwise error rate but generally has a higher type II error rate than REGWF for all pairwise comparisons

Alpha= 0.05 df= 20 MSE= 0.346112  
Critical Value of F= 2.86608  
Minimum Significant Difference= 1.2598

Means with the same letter are not significantly different.

Scheffe Grouping	Mean	N	VENDEDOR
A	9.0620	5	1
A	7.9840	5	4
B	5.7620	5	2
B	5.4820	5	3
B	4.9420	5	5

Si se tiene un modelo con más de un factor y se quiere hacer el contraste múltiple de los diferentes niveles de uno o unos factores, todo se haría lo mismo teniendo en cuenta, solamente, que habría que utilizar como término de error el mismo del factor que se esta descomponiendo.

Veamos como ejemplo de un factorial sin repetición el mismo visto en la página 182.

#### Archivo del programa SAS (C10-10.SAS).-

```
title 'Comparación múltiple de medias';
options ls=75 ps=60;
data media;
infile 'c10-10.dat';
input raza $ trata $ prod @@;
proc anova;
class raza trata ;
model prod = raza trata;
means trata / duncan lsd tukey snk scheffe;
run;
```

#### Archivo de datos (C10-10.DAT).-

```
T1 R1 5   T2 R1 7   T3 R1 4
T1 R2 1   T2 R2 1   T3 R2 0
T1 R3 4   T2 R3 6   T3 R3 3
T1 R4 2   T2 R4 2   T3 R4 1
```

#### Archivo de resultados (C10-10.LST).-

Dependent Variable: PROD					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	51.3333333	10.2666667	23.10	0.0007
Error	6	2.6666667	0.4444444		
Corrected Total	11	54.0000000			



R-Square	C.V.	Root MSE	PROD Mean
0.950617	22.22222	0.66667	3.00000

Source	DF	Anova SS	Mean Square	F Value	Pr > F
RAZA	2	8.0000000	4.0000000	9.00	0.0156
TRATA	3	43.3333333	14.4444444	32.50	0.0004

T tests (LSD) for variable: PROD

NOTE: This test controls the type I comparisonwise error rate not the experimentwise error rate.

Alpha= 0.05 df= 6 MSE= 0.444444

Critical Value of T= 2.45

Least Significant Difference= 1.3319

Means with the same letter are not significantly different.

T Grouping	Mean	N	TRATA
A	5.3333	3	R1
A	4.3333	3	R3
B	1.6667	3	R4
B	0.6667	3	R2

Duncan's Multiple Range Test for variable: PROD

NOTE: This test controls the type I comparisonwise error rate, not the experimentwise error rate

Alpha= 0.05 df= 6 MSE= 0.444444

Number of Means 2 3 4

Critical Range 1.332 1.380 1.404

Means with the same letter are not significantly different.

Duncan Grouping	Mean	N	TRATA
A	5.3333	3	R1
A	4.3333	3	R3
B	1.6667	3	R4
B	0.6667	3	R2

Student-Newman-Keuls test for variable: PROD

NOTE: This test controls the type I experimentwise error rate under the complete null hypothesis but not under partial null hypotheses.

Alpha= 0.05 df= 6 MSE= 0.444444

Number of Means 2 3 4

Critical Range 1.3319306 1.6700894 1.8843144

Means with the same letter are not significantly different.

SNK Grouping	Mean	N	TRATA
A	5.3333	3	R1
A	4.3333	3	R3
B	1.6667	3	R4
B	0.6667	3	R2

Tukey's Studentized Range (HSD) Test for variable: PROD

NOTE: This test controls the type I experimentwise error rate, but generally has a higher type II error rate than REGWQ.

Alpha= 0.05 df= 6 MSE= 0.444444

Critical Value of Studentized Range= 4.896

Minimum Significant Difference= 1.8843

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	TRATA
A	5.3333	3	R1
A	4.3333	3	R3
B	1.6667	3	R4
B	0.6667	3	R2

Scheffe's test for variable: PROD

NOTE: This test controls the type I experimentwise error rate but generally has a higher type II error rate than REGWF for all pairwise comparisons

Alpha= 0.05 df= 6 MSE= 0.444444

Critical Value of F= 4.75706

Minimum Significant Difference= 2.0563

Means with the same letter are not significantly different.

Scheffe Grouping	Mean	N	TRATA
A	5.3333	3	R1
A	4.3333	3	R3
B	1.6667	3	R4
B	0.6667	3	R2

Como se estudió, por primera vez, en los modelos factoriales con repetición, el término de error del factor en cuestión no tiene porque ser el *Error*, sino que puede ser una interacción u otro factor, en dicho caso, los contrastes múltiples de medias para los diferentes niveles de dicho factor tienen que hacerse con el cuadrado medio del término de error que le corresponda.

Veamos un ejemplo de un modelo mixto de dos factores con repetición en el que el término de error de los tratamientos es la interacción.

### Ejemplo.-

Se han probado cuatro tratamientos en cuatro rebaños elegidos al azar. Los cuatro tratamientos son cuatro dietas. Los datos son

		<i>Tratamientos</i>			
		$T_1$	$T_2$	$T_3$	$T_4$
$R_1$		73	82	45	55
		33	72	85	85
$R_2$		12	53	32	31
		22	83	22	51
$R_3$		69	64	34	64
		19	94	34	34
$R_4$		27	66	18	82
		87	76	48	72

## Archivo del programa SAS (C10-11.SAS)-

```

title 'Comparación múltiple de medias';
options ls=75 ps=60;
data medias;
infile 'c10-11.dat';
input rebano $ trata $ prod @@;
proc anova;
  class rebano trata ;
  model prod = rebano trata rebano*trata;
  test H=trata E=rebano*trata;
  means trata / duncan lsd tukey snk scheffe E=rebano*trata;
run;

```

## Archivo de datos (C10-11.DAT)-

```

R1 T1 73 R1 T2 82 R1 T3 45 R1 T4 55
R2 T1 12 R2 T2 53 R2 T3 32 R2 T4 31
R3 T1 69 R3 T2 64 R3 T3 34 R3 T4 64
R4 T1 27 R4 T2 66 R4 T3 18 R4 T4 82
R1 T1 33 R1 T2 72 R1 T3 85 R1 T4 85
R2 T1 22 R2 T2 83 R2 T3 22 R2 T4 51
R3 T1 19 R3 T2 94 R3 T3 34 R3 T4 34
R4 T1 87 R4 T2 76 R4 T3 48 R4 T4 72

```

## Archivo de resultados (C10-11.LST)-

Dependent Variable: PROD

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	11575.5000	771.7000	1.68	0.1570
Error	16	7350.0000	459.3750		
Corrected Total	31	18925.5000			

	R-Square	C.V.	Root MSE	PROD Mean
	0.611635	39.78290	21.4330	53.8750

Source	DF	Anova SS	Mean Square	F Value	Pr > F
REBANO	3	3476.50000	1158.83333	2.52	0.0945
TRATA	3	5977.50000	1992.50000	4.34	0.0204
REBANO*TRATA	9	2121.50000	235.72222	0.51	0.8443

Tests of Hypotheses using the Anova MS for REBANO\*TRATA as an error term

Source	DF	Anova SS	Mean Square	F Value	Pr > F
TRATA	3	5977.50000	1992.50000	8.45	0.0055

T tests (LSD) for variable: PROD

NOTE: This test controls the type I comparisonwise error rate not the experimentwise error rate.

Alpha= 0.05 df= 9 MSE= 235.7222

Critical Value of T= 2.26

Least Significant Difference= 17.366

Means with the same letter are not significantly different.

T	Grouping	Mean	N	TRATA
	A	73.750	8	T2
B	A	59.250	8	T4
B	C	42.750	8	T1
	C	39.750	8	T3

Duncan's Multiple Range Test for variable: PROD

NOTE: This test controls the type I comparisonwise error rate, not the experimentwise error rate

Alpha= 0.05 df= 9 MSE= 235.7222

Number of Means      2      3      4  
 Critical Range    17.37 18.13 18.56

Means with the same letter are not significantly different.

Duncan Grouping		Mean	N	TRATA
	A	73.750	8	T2
B	A	59.250	8	T4
B	C	42.750	8	T1
	C	39.750	8	T3

Student-Newman-Keuls test for variable: PROD

NOTE: This test controls the type I experimentwise error rate under the complete null hypothesis but not under partial null hypotheses.

Alpha= 0.05 df= 9 MSE= 235.7222

Number of Means      2      3      4  
 Critical Range    17.365114 21.433208 23.964921

Means with the same letter are not significantly different.

SNK Grouping		Mean	N	TRATA
	A	73.750	8	T2
B	A	59.250	8	T4
B		42.750	8	T1
B		39.750	8	T3

Tukey's Studentized Range (HSD) Test for variable: PROD

NOTE: This test controls the type I experimentwise error rate, but generally has a higher type II error rate than REGWQ.

Alpha= 0.05 df= 9 MSE= 235.7222  
 Critical Value of Studentized Range= 4.415  
 Minimum Significant Difference= 23.965

Means with the same letter are not significantly different.

Tukey Grouping		Mean	N	TRATA
	A	73.750	8	T2
B	A	59.250	8	T4
B		42.750	8	T1
B		39.750	8	T3

Scheffe's test for variable: PROD

NOTE: This test controls the type I experimentwise error rate but generally has a higher type II error rate than REGWF for all pairwise comparisons

Alpha= 0.05 df= 9 MSE= 235.7222  
 Critical Value of F= 3.86255  
 Minimum Significant Difference= 26.132

Means with the same letter are not significantly different.

Scheffe Grouping	Mean	N	TRATA
A	73.750	8	T2
B	59.250	8	T4
B	42.750	8	T1
B	39.750	8	T3

**Bibliografía:**

- Cochran, WG y Cox, GM.* 1965. DISEÑOS EXPERIMENTALES. Ed. TRILLAS. México.
- Li, Ching Chun.* 1969. INTRODUCCIÓN A LA ESTADÍSTICA EXPERIMENTAL. Ed. Omega. Barcelona.
- Lite, TM, y Jackson Hills, F.* 1987. MÉTODOS ESTADÍSTICOS PARA LA INVESTIGACIÓN EN LA AGRICULTURA. Ed. TRILLAS. México.
- Littel, R.C., Freund, R.J. and Spector, P.C.* 1991. SAS® FOR LINEAR MODELS. SAS Institute Inc., Cary, NC, USA.
- Ostle, B.* 1965. ESTADÍSTICA APLICADA. Ed. Limusa-Wiley. México.
- Ruiz-Maya, L.* 1983. MÉTODOS ESTADÍSTICOS DE INVESTIGACIÓN. Ed. INE. Madrid.
- Sokal, R.R. y Rohlf, F.J.* 1994. BIOMETRY. Ed. W.H.FREEMAN. San Francisco.
- Snedecor, G.W. y Cochran, W.G.* 1971. MÉTODOS ESTADÍSTICOS. Ed C.E.C.S.A. México.
- Steel, R.* 1996. PRINCIPLES AND PROCEDURES OF STATISTICS. Ed. McGRAW-HILL Education. New York .
- SAS Institute Inc.* 1990. SAS/STAT USER'S GUIDE. Volume 1 and 2. Cary, NC, USA.